COMPARATIVE GENOMICS OF THE MICROBIAL CHEMOTAXIS SYSTEM

A Dissertation
Presented to
The Academic Faculty

by

Kristin Wuichet

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in Biology

Georgia Institute of Technology
August, 2007

COMPARATIVE GENOMICS OF THE MICROBIAL CHEMOTAXIS SYSTEM

Approved by:

Dr. Igor Zhulin, Advisor
School of Biology
University of Tennesee – Oak Ridge National Laboratory

Dr. King Jordan
School of Biology
Georgia Institute of Technology

Dr. Mark Borodovsky
School of Biology
Georgia Institute of Technology

Dr. Gladys Alexandre
Department of Microbiology
University of Tennessee

Dr. Stephen Harvey
School of Biology
Georgia Institute of Technology

Date Approved: May, 17, 2007

ACKNOWLEDGEMENTS

I want to thank Dr. Igor Zhulin for encouraging me to pursue bioinformatics when I came to him as a curious experimental scientist and helping me to understand the bigger picture of science and life. I want to thank my committee members, Dr. Gladys Alexander, Dr. Mark Borodovsky, Dr. Stephen Harvey, and Dr. King Jordan, for their helpful participation. I also want to thank my mother and stepfather for continuing to support me no matter the situation, and my father for constantly asking me when I will get a "real job." Ben, your patience has been tested through this time, but your stability and humor helped me to keep me sane. Shelby, you were always there for me when I needed a beer buddy or just someone to listen to me whine. Roger, collaborating with you in the bowels of Cherry Emerson sparked wonderful ideas that led to much of this work. Luke, you always came through for me when I needed a program to run or high-throughput analysis beyond my capabilities, and you inspired me to push my own programming boundaries.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

24-64H	24–64 Heptad
a.a.	amino acid
Alt	Alternative output
BLAST	Basic Local Alignment and Search Tool
CheA-Alt	Alternative output system associated CheA
CheA-Fla	Flagellar system associated CheA
CheA-Tfp	Type IV pili system associated CheA
CheY-P	Phosphorylated CheY
DNA	Deoxyribonucleic Acid
F1-10	Flagellar chemotaxis system classes 1-10
GI	Genbank Identifier
HKI	Class I Histidine Kinase
HKII	Class II Histidine Kinase
HMM	Hidden Markov Model
Hpt	Histidine phosphotransfer
MCP	Methyl-accepting Chemotaxis Protein
ME	Minimum Evolution
MEGA	Molecular Evolutionary Genetics Analysis
MI	MiST Identifier
MiST	Microbial Signal Transduction
NCBI	National Center for Biotechnology Institute

NJ	Neighbor-Joining
OCS	One-Component System
PDB	Protein Data Bank
Pfam	Protein families
PSI-BLAST	Position Specific Iterative Basic Local Alignment and Search Tool
REC	Receiver
RefSeq	Reference Sequence
rRNA	Ribosomal Ribonucleic Acid
SMART	Simple Modular Architecture Research Tool
TCS	Two-Component System
Tfp	Type IV pili
tRNA	Transfer Ribonucleic Acid

SUMMARY

This research project presents a comprehensive functional analysis of a complex prokaryotic signal transduction system and the mechanisms underlying its evolution. Signal transduction systems are protein interaction networks that govern complex biological processes that are essential to many organisms like metabolism, biosynthesis, motility, and virulence. The chemotaxis system regulates motility in prokaryotes and is their most complex signal transduction system. The system has been extensively characterized experimentally, but recent studies have created new questions about the function and origin of this system. Experimental research cannot keep up with questions generated by the exponentially growing amounts of genomic data. Instead computational methods can be used to help focus future research and extrapolate current experimental data onto the genomic landscape. Comparative genomics analyses are well-suited for studying the chemotaxis system since it is present in taxonomically diverse organisms.

The first aim of this project is to understand the evolutionary history of the chemotaxis system that has resulted in the diversity of chemotaxis systems that have been experimentally. The results reveal three functional families of chemotaxis systems that regulate flagellar motility, type IV pili motility, and non-motility outputs. The flagellar family shows extensive diversity with 10 conserved classes that have variable accessory proteins, and these classes show a co-evolutionary relationship with flagella. The chemotaxis system is predicted to have evolved from two-component signal transduction systems. Hybrid systems that represent a functional link between two component and chemotaxis systems were identified, and they may also be an evolutionary link between

the two systems. The results of this aim provide new insights about the evolution of this system as well as general paradigms of prokaryotic evolution and functional information for experimental scientists about the specific interactions involved in each chemotaxis system family and class.

The second aim of this project is to analyze the molecular evolution of chemotaxis system components and utilize that information to predict the contact sites involved in protein-protein interactions. The first part of the analysis provides empirical evidence that sensory domains of proteins evolve faster at the amino acid level than protein signaling domains. The data supports that the slower evolution of signaling domains is due to their involvement in one or more protein-protein interactions, unlike many signaling domains. Since protein-protein interactions are under evolutionary pressure to be maintained, the sequence conservation between core components of a chemotaxis system that also interact with accessory components and homologous core components that lack the accessory interaction were examined for differences. Residues specifically conserved in the accessory interaction set are involved in maintaining the interaction based on experimental validation and mapped to protein-protein interfaces of available three-dimensional structures. The results show that the key residues involved in maintaining an interaction can be predicted based on sequence information alone, which will be of value to experimental scientists studying protein-protein interactions.

CHAPTER 1

INTRODUCTION

The availability of completely sequenced genomes has undergone an explosion since the advent of whole-genome shotgun sequencing [1]. There are currently almost 500 completely sequenced prokaryotic and eukaryotic genomes with nearly 2000 more on the way. The motivation for sequencing genomes comes with the implication that new biological insights can be derived from it; however, a significant decline in high-impact journal publications per number of sequenced genomes shows that novel conclusions cannot be gleaned from genomes without more rigorous comparative analysis [2]. There are a variety of goals for comparative genomics, such as: (i) evaluating the differences between bacteria in search for determinants of virulence [3-5], antibiotic resistance, metabolic properties, etc, (ii) analysis of genomes and individual genes to understand the mechanisms of evolution [6,7], and (iii) comparison of homologous proteins and systems across the genomic landscape to reveal novel functions and mechanisms [8,9]. The aim of the research presented in this thesis falls into the latter two categories with a focus on a comprehensive functional analysis of a complex prokaryotic signal transduction system and the mechanisms underlying its evolution.

Signal transduction systems are present in all organisms in order to sense internal and external cues and trigger appropriate cellular responses. Prokaryotic signal transduction can be classified into three main categories: one-component systems (OCSs) [10], two-component systems (TCSs) [11-14], and the chemotaxis system [15-17]. As their name suggests, one-component systems are made up of a single protein that is capable of both sensing a signal and directly affecting a cellular response, either through a single domain (such as a DNA binding domain that acts through a metal cofactor) or multiple domains (separate input and output domains). Due to their single protein nature

and typical lack of transmembrane spanning regions, one-component systems are predicted to primarily sense the internal cellular environment. Separation of input and output between two or more proteins in two-component systems enables both internal and external signals, as most TCS sensors are predicted to be membrane-bound, and contain intracellular and extracellular sensory domains. Like TCSs, the chemotaxis system also consists of multiple proteins separating input and output and utilizes similar sensing and signaling modules. Protein components of TCSs and chemotaxis systems have the same type of ATP-binding and phosphoreceiver domains. There are two primary differences between TCSs and chemotaxis systems. The first one is in their molecular design, where the TCS network in the cell comprises parallel systems that link multiple inputs to multiple outputs, such as various promoter sites (Figure 1.1). In contrast, the chemotaxis system is organized in such a way that it links multiple inputs (similar to those detected by TCSs) to a single output – a motility organelle (Figure 1.1). Second, additional system components lacking from TCSs confer unique temporal sensing capabilities upon the chemotaxis system. Other noticeable differences include clustering of the chemotaxis signaling complex at the cell pole (which is similar to our own olfactory systems that utilize a discrete signaling patch) and a form of molecular memory, which is missing from simple TCSs. In the evolutionary perspective, one-component systems are proposed to be the most ancient form of signal transduction, because they are simple in molecular design, show the greatest domain diversity among all classes of signal transduction and are found in all kingdoms of life. On the other hand, two component systems are thought to have originated later in evolution, within bacteria, and spread to some archaea and very few eucarya by horizontal transfer.

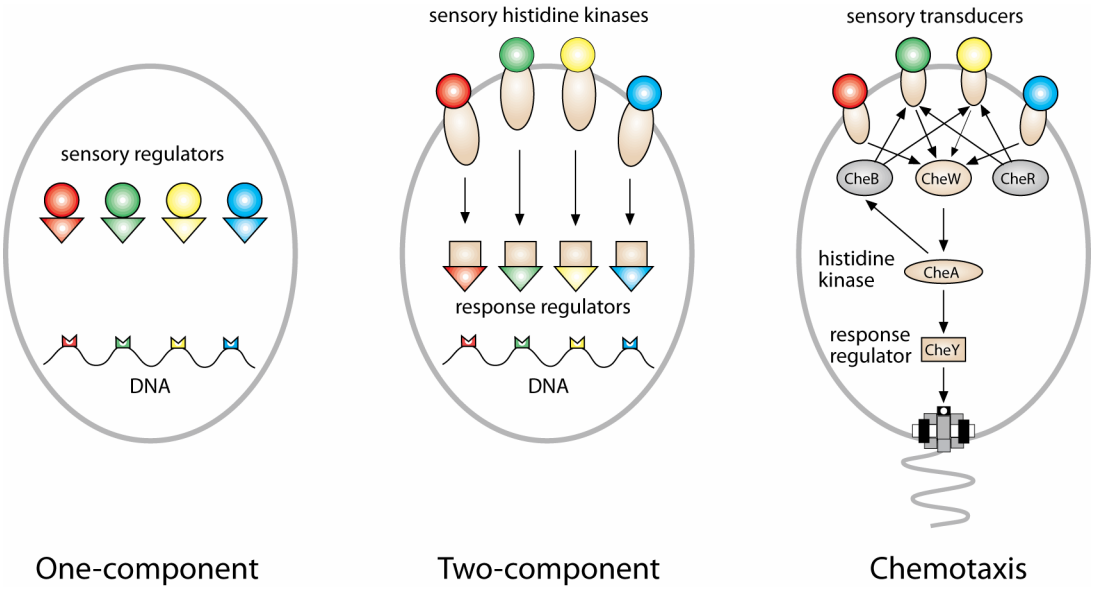


Figure 1.1 Paradigms of prokaryotic signal transduction. Circles and triangles represent dedicated input and output domains. Ovals and rectangles represent transmitting and receiving modules of signal transduction. Gray ovals represent receptor modifying enzymes of the chemotaxis system.

We have chosen to apply computational genomic methods to study the most complex and best characterized signal transduction system of prokaryotes, the chemotaxis system. Motivating factors for the decision to study the chemotaxis system include: (1) the wealth of genomic data since 477 of the 497 completely sequenced genomes available to date are from prokaryotes; (2) the large evolutionary distances between prokaryotes that have this system, which is important for comparative analysis; (3) the propensity for interacting chemotaxis proteins to be encoded in gene clusters, which enables productive gene neighborhood analysis; (4) extensive genetic and biochemical characterization of the system and its components [15-17] that establishes solid references for functional predictions; (5) the availability of three-dimensional structures for almost all of the components [18-32], which provides a framework for functional assignments; (6) questions about the function and origin of this system that cannot be answered by experimental methods [33,34]. The research design and methods

presented here would be applicable to other multi-component systems in prokaryotes and eukaryotes,

The chemotaxis system is a network of interacting proteins, which senses environmental signals to regulate flagellar motility. The system consists of two distinct pathways, an excitation pathway that has the downstream result of interacting with the flagellum and an adaptation pathway that provides a temporal response for sensing the signal gradient. The excitation pathway is composed of methyl-accepting chemotaxis proteins (MCPs) that detect environmental signals and transmit them via their highly conserved cytoplasmic signaling domain to a scaffolding protein, CheW, and a histidine kinase, CheA [15-17]. The signals regulate the kinase activity of CheA for its receiver and final core component, CheY, and the phosphorylation state of CheY controls its affinity for the flagellar motor. In some organisms, one or more phosphatases (CheC, CheX, and/or CheZ) are involved in the secondary adaptation pathway that aid in dephosphorylating CheY [33]. The signal propagation through the MCPs is further controlled by the primary adaptation pathway: the CheR methyltransferase that constitutively methylates particular MCP residues and the CheB methylesterase, a response regulator that is phosphorylated by CheA to stimulate the removal of methyl groups from these residues [33]. Many chemotaxis systems have an additional protein, CheD, an enzyme that carries out deamidation of amino acid side chains prior to their methylation by CheR [35]. In some of these systems, the MCP-CheD interaction is dependent on a CheD-CheC interaction [29,36]. The final protein known to play a direct role in the chemotaxis system is CheV, a fusion of the CheW scaffolding protein and the CheY response regulator that affects the signaling state of the MCP based on its phosphorylation state as controlled by the CheA kinase [37]. The MCPs, CheW, and CheA proteins form a large complex [38] that is located in the cytoplasmic membrane at one or both poles of the cell [39-41] or sometimes even in the cytoplasm [42,43], and,

along with CheY, comprise the core set of components that are required for a functional system.

It is fundamentally important in any protein analysis, experimental or computational, to recognize the domain nature of protein architecture. Domains are defined as conserved elements in sequence (usually ~ 100 amino acid long regions showing conservation of secondary structure elements and functional motifs), structure (a smallest unit of the protein that folds independently), and overall function. Conserved sequence features allow for the identification of domains within any given sequence by Hidden Markov Model (HMM) analysis [44,45]. HMMs are built from multiple alignments of domain families found through similarity searches such as the BLAST (Basic Local Alignment Search Tool) algorithm [46] of sequence repository databases like NCBI and EMBL. The domain architectures of members of a protein family can reveal which domains are core elements that are essential for protein function or accessory elements that may confer additional functions in certain family members. Furthermore, the use of core domains is essential for proper phylogenetic analysis in order to remove biases caused by accessory domains not present in all proteins. Lastly, domain architecture can be used for protein-protein interaction analysis since certain domains are known to be involved in specific interactions based on biochemical and structural studies [11].

In addition to providing information about protein interactions, modifications, functions, and structures, experimental studies have also shown that there can be multiple chemotaxis pathways within a single organism [41,47-49]. Some of these systems regulate non-flagellar motility [50], and even more surprisingly, they can have an output entirely unrelated to motility [51-53], such as regulation of gene expression [54]. These findings give increased importance to finding ways of distinguishing the primary interacting protein partners when paralogous and/or horizontally transferred systems are present. Specifically, it is important to identify which systems may have a non-motility

related output, and to analyze whether there is a potential cross-talk between paralogous systems. Although domain architecture can reveal multiple proteins that might interact, gene neighborhood analysis provides further valuable information, because interacting proteins are often encoded in close proximity on a chromosome in prokaryotic genomes. Indeed, many chemotaxis systems are encoded in operons and compact gene clusters [48,49,51,53-65]. Additionally, gene neighborhood can be used to identify potential alternative output response regulators when the primary chemotaxis output protein, CheY, cannot be found for a particular system. In order to evaluate the potential for cross-talk between chemotaxis proteins in organisms that have paralogous and horizontally transferred components and/or systems, more information is needed about the details of the protein-protein interactions, e.g. contact sites that govern the interaction specificities.

Protein-protein interactions are essential for mediating signal propagation in complex systems, such as signal transduction and metabolic pathways [66]. There has been a significant effort in both experimental and computational communities towards the development of techniques that can be used to predict biologically relevant protein-protein interactions. Experimental studies have focused on the high throughput application of standard interaction analyses, such as the yeast two-hybrid system [67], co-immunoprecipitation [68], cross-linking [69], and affinity chromatography [70], as well as newer technologies, such as protein microarrays [71]. Computational methods utilize domain architecture [72,73], phylogenetic [74-76], gene neighborhood [77,78], and/or co-crystal information in order to predict potential protein-protein interactions with varying levels of confidence. Understanding protein interactions in cell regulatory networks is of great importance. In order to obtain in-depth knowledge of the actual mechanisms of signal propagation, more information on the specific protein-protein contact sites is required. Co-crystal structures of interacting partners provide valuable insight about the nature of protein-protein interfaces, but (i) there is a lack of predictive methods for

evaluating potential contact sites in interacting protein partners that lack co-crystals and (ii) co-crystals usually capture a single rigid state of an interaction, which may not fully reflect its dynamics. Interacting amino acid residues within a protein and between two proteins can also be found using experimental mutagenesis, such as second site suppressor analysis [79-81]. Structural and experimental data about particular protein interaction interfaces can be used to guide and verify computational predictions derived from the wealth of genomic data leading to the development of novel, predominantly computational approaches to predict contact sites in protein-protein interactions.

Proteins that interact co-evolve [82], which is reflected in the evolutionary analysis of both members of an interaction [74]. These observations and experimental second site suppressor studies laid a foundation for computational approaches that strive to find coordinating changes between amino acids in families of homologous proteins, which could be the sites of intraprotein [83,84] or interprotein interactions [85,86]. All of the methods involve creating multiple sequence alignments of protein sets whose members are predicted to interact with each other. The resulting alignments are then analyzed to find correlated changes in particular sequence positions or used for building phylogenetic trees that help to group the sequences into related subfamilies as seen by organism distribution. Many computational methods are developed with the intent to further automate them so that predictive information can be rapidly produced creating potential targets for experimental validation. In the field of protein interaction prediction, there are still many reasons for detailed manual analysis rather than automation because automated methods do not yet have the flexibility to handle different levels of sequence conservation found in protein-protein interactions nor is the nature of amino acid substitutions in maintaining these interactions well understand.

The aim of this research project is to perform a computational analysis of the chemotaxis system at several levels of its organization starting from understanding the diversity of its components and working towards deciphering the local protein-protein

interactions between these components at the amino acid level. The fundamental theory underlying this study is evolution and how it manifests at the organismal, domain, and molecular level. First, there is an analysis of the presence and absence of the components of the system in order to infer its evolutionary history. Phylogenetic and gene neighborhood analyses of the system components are considered together with the wealth of available experimental data to link distinct subfamilies of chemotaxis systems to their functional properties. Then we present a case study of highly conserved chemotaxis operons in cyanobacteria to examine the rates of evolution of different domains within the only components of the chemotaxis system that have both extracellular and intracellular modules - MCPs. Finally, rigorous amino acid sequence analysis of subfamilies of interacting chemotaxis proteins is performed to identify probable interacting proteins, with additional support from gene neighborhood information, and to find the selectively conserved residues that are potential protein-protein contact sites, with verification using structural and experimental data. To our knowledge, this type of multi-level analysis has not been done for any other signal transduction, metabolic, or other multi-protein system and will provide a framework for future computational studies of complex systems.

CHAPTER 2

GENERAL SOURCES, TOOLS, AND METHODS

2.1 Database Sources

2.1.1 Sequence databases

The Microbial Signal Transduction (MiST) [87] and Reference Sequence (RefSeq) [88] databases are the sources of all sequence data used in this research. The RefSeq database is a non-redundant set of sequences from draft and complete genomes maintained by the National Center for Biotechnology Information (NCBI) and built from its GenBank [89] database. The GenBank database is a comprehensive public repository of nucleotide sequences. In RefSeq proteins as well as tRNAs and rRNAs are annotated and directly linked to the nucleotide information. Proteins are annotated as a whole, and conserved functional regions within the proteins are further described. Literature relevant to particular proteins is also linked. Format consistency allows easy retrieval of sequences or sequence regions. Each sequence is given a stable accession number and a unique GenBank Identifier (GI) separate from its GenBank database information. The RefSeq database has ongoing curation to continually reflect current knowledge of biological and sequence data.

The MiST database is built from the complete genomes of the NCBI RefSeq database. It includes the full nucleotide and a translated protein set along with the full annotation of each gene and encoded protein from RefSeq. Each protein is associated with its RefSeq GI number as well as a unique MiST identification number. Each protein sequence is scanned against the domain models from the Pfam-A [90] and SMART [91] databases, and a graphic image for each protein is provided that reveals conserved

domains (see more in Chapter 2.1.2) and predicted low complexity, transmembrane, coiled-coil, and signal peptide regions. Additionally, the amino acid and associated nucleotide sequence for each protein are provided, along with a graphic representation of neighboring genes that can be individually accessed to reveal the information associated with its encoded protein. A signal transduction profile is provided for each genome, which shows the numbers of various OCS and TCS proteins. Individual and multiple genomes can be queried by GI or MiST identification numbers, domain model names, or a general description, such as a protein name. Only data from completely sequenced genomes available from RefSeq as of July 1, 2006 or earlier were used for this study. Redundant species were removed from final analysis resulting in a final genome set shown in Table A.1.

2.1.2 Domain databases

Protein sequences can share similarity across their entire lengths or within discrete regions. In 1973 Wetlaufer defined domains as distinct structural regions within a protein that are capable of folding autonomously [92]. Since then scientists have found that many structural domains can be identified by sequence similarities within a domain family [93-95]. A protein may consist of a single domain or multiple domains that can each confer a different function in the protein. Hidden Markov Models (HMMs) [44,45] can be built from multiple sequence alignments of protein domains. HMMs capture the key features common to members of a multiple alignment and can be used to identify homologous regions in a new sequence without building a new multiple alignment. The ability of an HMM to identify homologous regions is dependent on the quality of the multiple alignment from which it was built. Ideally the multiple alignment will contain all known members of a protein family or subfamily, or a representative sample. HMMs built from alignments that lack a representative subset can fail to identify more divergent

members of the sequence family. All of the HMMs used to identify protein domain architecture come from the Pfam [90] and SMART [91] domain databases.

The Pfam database strives to encompass all protein domains and version 21.0 released in November, 2006 contains 8957 curated domain models in Pfam-A (http://www.sanger.ac.uk/Software/Pfam/). Pfam also contains a set of automatically generated domain models (Pfam-B) based on motifs identified in the ProDom [96] database. Often multiple domains share a common evolutionary origin, and thus some distant sequence similarity. Pfam does not allow a sequence region from one domain alignment to be a member of other domain alignments. When a protein sequence is homologous to multiple members of a superfamily of domains, the highest scoring domain is chosen and identified as a member of a clan of related domain models. The Pfam database can be accessed by a web server where graphical domain architecture for any given sequence are provided, which includes Pfam domain, signal peptide, coiled-coil, low complexity, and transmembrane regions, as well as detailed information for each domain model including its function, taxonomic distribution, relevant literature, domain architectures of the proteins of all domain family members, and structural information if available. Pfam domain models also can be downloaded freely for local use.

The SMART database [91] is significantly smaller than Pfam with 726 curated domain models currently, but it is focused specifically on difficult-to-identify, genetically mobile domains, particularly those involved in signal transduction. Like Pfam, the SMART web server provides graphical domain architectures that include domain, signal peptide, coiled-coil, low complexity, and transmembrane regions, but it uses both SMART and Pfam-A domain models. The SMART 5.0 release also provides a prediction of the catalytic activity of 50 of its domain models, as well as protein interaction information imported from the STRING [73] database. Information related to a domain's function, structure, and taxonomic distribution, as well as published literature

on it is also available for each SMART domain model. SMART domain models can also be downloaded freely.

2.1.3 Structural database

The Protein Data Bank [97] is the worldwide repository of three-dimensional structure data for proteins and nucleic acids. Chemotaxis components with available three-dimensional structures (individual and co-crystals) used for analysis were retrieved from the PDB, CheA: 1B3Q, 1I5N; CheA/CheY: 1U0S, 1EAY; CheY/CheZ: 1KMI, CheC/CheD: 2F9Z.

2.2 Tools

2.2.1 Sequence Similarity

2.2.1.1 Basic Local Alignment and Search Tool (BLAST)

BLAST [46] enables quick identification of protein homologs by finding local regions of similarity between a given sequence and sequences in a selected database. NCBI maintains BLAST allows searches of against any of its sequence databases as well as the PDB and SWISS-Prot database [98]. The genomic BLAST pages are broken up into sections such as microbial and eukaryotic where genomes within each set can be selected individually or in groups for more tailored results. BLAST enables protein-protein, DNA-DNA, DNA-protein, and protein-DNA sequence comparisons. With the latter two, the DNA sequences are translated into amino acid sequences. BLAST works by identifying at least two short non-overlapping regions (known as "words") between two sequences that are of high similarity and in close proximity to each other and then extending the sequence comparison between and beyond the words. A score is assigned based on the amount of similarity between the two sequences and the length of the region of comparison. The length of the region that results in the best score is used, and

sequences with scores above a given threshold are returned from the search. BLAST has numerous options for users such as multiple substitution matrices, gap parameters, and low complexity filters for the sequence comparisons.

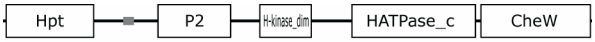
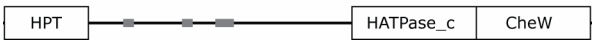
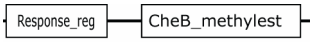
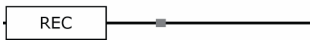
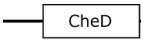
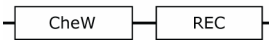
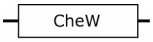
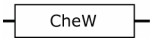
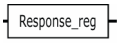
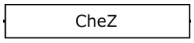
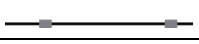
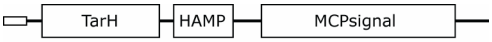
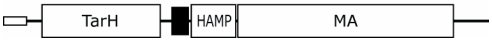
2.2.1.2 Position Specific Iterative BLAST (PSI-BLAST)

PSI-BLAST [46] takes the results of an initial BLAST search and builds a multiple alignment of all of the significant sequences. A Position Specific Score Matrix (PSSM) is built from the multiple alignment, which, similar to an HMM, captures the likelihood of finding a given character at certain positions in the alignment. The next round of BLAST uses the PSSM instead of the initial substitution matrix. Given that homologous sequences are often conserved only in discrete regions, the PSSM allows the BLAST program to better identify distant homologs. The PSSM is updated with each round of PSI-BLAST to include all significant sequences. Once no new homologues are identified the search is said to have converged. To identify more divergent sequences the threshold for inclusion in the PSSM can be lowered, and sequences below threshold that are suspected to be homologs can be manually included.

2.2.2 Domain Architecture Analysis

All protein domain architectures were initially determined using the pre-computed results in the MiST database [87]. MiST uses the HMMER software package [99] to identify the domain architecture of proteins using the domain models from Pfam and SMART. Large stretches (>75 amino acids, a.a.) in protein sequences that lacked any identifiable domains and coiled-coil or low complexity regions were subjected to PSI-BLAST searches in order to identify new domains, or reveal known domains that have not been identified by current domain models. In this work, we used a local PSI-BLAST web server at the Oak Ridge National Laboratory (ORNL), which provides several new options for PSI-BLAST output including formatting, sequence retrieval and graphical

Table 2.1 Domain architecture of chemotaxis proteins as visualized in MiST. The MiST database uses the domain models from both Pfam and SMART databases. Domains are shown as white boxes with their names inside. Small black, gray, and white boxes indicate predicted coiled-coil, low complexity, and signal peptide regions, respectively. Large black boxes indicate predicted transmembrane regions. The NCBI database GI numbers corresponding to each protein sequence are given under their respective protein names.

Protein GI	Database	Domain Architecture
CheA 15643465	Pfam	
	SMART	
CheB 15802295	Pfam	
	SMART	
CheC 15643666	Pfam	
	SMART	
CheX 15644366	Pfam	
	SMART	
CheD 15643665	Pfam	
	SMART	
CheR 15802296	Pfam	
	SMART	
CheV 16078465	Pfam	
	SMART	
CheW 15802299	Pfam	
	SMART	
CheY 15802294	Pfam	
	SMART	
CheZ 15802293	Pfam	
	SMART	
MCP 15802298	Pfam	
	SMART	

improvements (Luke Ulrich, unpublished data). Domain architectures of chemotaxis proteins used in this study are shown in Table 2.1.

2.2.3 Gene Neighborhood Identification

Gene neighborhoods for proteins were determined by analyzing neighboring proteins in the MiST database in either direction until six or more consecutive non-chemotaxis proteins (i.e. any protein other than CheA, CheB, CheC, CheD, CheR, CheV, CheW, CheX, CheY, CheZ, or MCP) were identified. In order to quickly identify whether or not two proteins were in the same gene neighborhood we took advantage of the consistent numbering scheme for consecutive encoded proteins in the MiST database. Although adjacently encoded proteins have consecutively numbered GIs in the majority of RefSeq genomes, a few genomes of interest (*Burkholderia thailandensis E264* [100], *Carboxydothermus hydrogenoformans Z-2901* [101], and *Myxococcus xanthus DK 1622* [102]) lack a consistent numbering scheme, and other genomes had proteins that were identified and annotated after their official release, which also resulted in inconsistent numbering in discrete regions. The MiST protein numbering scheme is entirely sequential, which allowed quick gene neighborhood labeling through the use of PERL scripts that were written to link together the chemotaxis proteins collected for this study based upon their MiST Identifier (MI) number.

2.2.4 Multiple Sequence Alignment

TheClustalX v1.83 [103] and MUSCLE v3.6 [104] software packages were used to build the multiple sequence alignments used in this study, and subsequent manual editing was done using SeaView [105]. The ClustalW algorithm [106] with default settings in ClustalX was used for initial multiple alignments of all protein sequence families families, except CheY, because of its ability to identify shared core regions

among homologous proteins despite domain architecture differences or significant sequence divergence with high speed and nominal gap insertion. MUSCLE was used for the alignment of CheY sequences since it is better equipped to handle large data sets and large domain architecture differences were not expected (see Methods 2.3.1). Conserved core regions of the CheA, CheB, and CheR multiple alignments were realigned with MUSCLE. Bacterial and Archaeal 16S sequence sets were separately aligned in MUSCLE because of their significant differences, and then the resulting two alignments were merged together using the profile-profile alignment feature of MUSCLE.

2.2.5 Phylogenetic Analysis

The MEGA v3.1 [107] and PHYML [108] software packages were used for the phylogenetic analyses. PHYML was used for maximum likelihood phylogenetic analyses. Neighbor-joining (NJ) and minimum evolution (ME) trees were built using MEGA . NJ and ME trees built from the same alignment often have similar topologies, but the ME method is more exhaustive in the search for the correct topology. MEGA uses the closest-neighbor interchange method to examine local differences in topology of multiple NJ trees and identify the ME tree. All trees were validated by system association predictions (Chapter 2.3.3) to ensure the optimum tree was selected. To eliminate potential biases, all trees were built from conserved core regions made up of domains common to all protein family members. The legends of Tables A.1, A.2-A.12, and A14-A16 detail the core regions used in phylogenetic analysis, as well as the methods used for building the optimal trees of each data set.

2.2.6 Secondary Structure Prediction

Individual secondary structure analysis was obtained using the Jnet prediction method [109] on the Jpred secondary structure prediction server (http://www.compbio.dundee.ac.uk/~www-jpred/). Predicted secondary structure of

multiple sequence alignment members was visualized using the VISSA program [110] that uses the PSIPRED prediction method [111] for each sequence.

2.2.7 Three-Dimensional Structure Visualization

The PyMol software package (http://www.pymol.org) was used for all three-dimensional structure visualizations.

2.2.8 Solvent Accessibility Prediction

Solvent accessibility of particular proteins was predicted using the the Jnet prediction method [109] on the Jpred secondary structure prediction server (http://www.compbio.dundee.ac.uk/~www-jpred/).

2.2.9 Pairwise Alignment

Pairwise alignment scores for amino acid sequences were determined by William Pearson's LALIGN program as implemented in the European Molecular Biology Network server (http://www.ch.embnet.org/software/LALIGN_form.html), using the global alignment without end gap penalties method and the default scoring matrix and gap penalty parameters (BLOSUM50, -14 opening gap , and -4 extending gap).

2.2.10 Sequence Conservation Analysis

CONSENSUS and WebLogo were used for sequence conservation analysis. CONSENSUS was used for domain analysis, and the WebLogo [112] webserver (http://weblogo.berkeley.edu/) was used to obtain sequence logo images that are better suited for quick visualization of smaller motifs. The CONSENSUS script was downloaded from the European Molecular Biology Laboratory (http://coot.embl.de/Alignment/Script/consensus.txt) and implemented locally (Luke Ulrich, unpublished work). The script parameters were changed as needed to identify specific physicochemical conservation levels.

2.3 Methods

2.3.1 16S rRNA Homolog Retrieval

16s rRNA sequences were used for subsequnet phylogenetic analyses in order to trace the evolution of the chemotaxis system based on its presence in major taxonomic clades. Given the differences between archaeal and bacterial 16S rRNA, annotated 16S sequences from *Halobacterium sp. NRC-1* and *Escherichia coli* were used as queries in PSI-BLAST searches in the microbial section of genomic BLAST against completely sequenced archaeal and bacterial genomes, respectively. The high conservation of 16S rRNA resulted in quick convergence in both searches. Hits that corresponded to annotated 16S sequences within the genomes were identified, and regions corresponding to the annotated 16S sequence plus 160 bases extended from the 3' and 5' ends were retrieved from the associated genomes. The hits were extended to ensure that the entire 16S region was captured for the best multiple sequence alignment and subsequent phylogenetic analysis. Introns within the 16S region of *Aeropyrum pernix K1* [113] resulted in poor annotation of the 16S region, and the 16S sequence was identified instead by the entire length of the BLAST hit regions with the standard 160 base extensions. The core 16S rRNA regions for each sequence as identified in the final multiple alignment (see Methods 2.2.3) and are given in Table A.1, which also shows the final genome set analyzed after redundant species were removed.

2.3.2 Protein Homolog Retrieval

In order to perform a comprehensive analysis of the chemotaxis system we aimed to retreive all members of each component protein family for subsequent multiple alignments and phylogenetic analyses. Protein-protein PSI-BLAST searches against the RefSeq database were conducted using full or partial sequences experimentally characterized CheA, CheB, CheR, CheW, CheZ, and FlhA proteins from *E. coli*; CheC,

CheD, and CheV proteins from *Bacillus subtilis*; CheX from *Thermotoga maritima*; PilU from *Pseudomonas aeruginosa*; and FlaH from *Halobacterium* sp. NRC-1 (all sequences can be found in Tables A.1-A.16). Standard PSI-BLAST parameters were used and sequences were collected once convergence was reached and sequences below threshold could reasonably be assumed not to be divergent homologs based on domain architecture and/or phyletic distribution. Homologs of many of these chemotaxis proteins were identified in the genome of the eukaryote, *Anopheles gambiae* [114]. It is presumed that these proteins are the result of a contamination, and they were removed from analysis since all of the remaining proteins are found in prokaryotic genomes. Only homologs from the genomes used in Table A.1 were retrieved for final analysis. All protein sets were confirmed by domain architecture analysis and/or multiple sequence alignments. The final proteins sets, their gene neighborhoods, and regions used for subsequent phylogenetic analyses are listed in Tables A.1-A.16.

2.3.2.1 CheB, CheD, CheR, and CheZ

Only the region corresponding to the enzymatic portion of CheB (Pfam:CheB_methylest) was used in a PSI-BLAST query in order to eliminate hits with its receiver domain (Pfam, SMART: REC) that is found in kinases and response regulators of many TCSs [115]. Full length sequences were used in the remaining PSI-BLAST queries, but the CheZ search was filtered to eliminate hits from sequences longer than 300 a.a. (the *E. coli* CheZ is 214 a.a.) due to similarity between the CheZ four helix coiled-coil region and similar, but longer, region of MCPs [23,25,31]. 92 out of 98 CheD proteins and 275 out of 280 CheB proteins were identified by the CheD and CheB_methylest domain models, respectively; multiple alignments and gene neighborhood analysis supported their assignment to chemotaxis pathways. Domain

queries against MiST for proteins with the CheB_methylest or CheD domains did not reveal any proteins that were not previously identified by sequence similarity searches. Domain architecture analysis revealed that all CheR proteins contain the Pfam:CheR and SMART:MeTrc domains, and domain queries for CheR and MeTrc against the MiST database identified the same sequence set as the similarity searches. Domain architecture queries for CheR or MeTrc picked up additional proteins due to distant homology to other methyltransferase proteins better described by other domain models. Only 49 of the 75 CheZ proteins retrieved by similarity searches were identified by the CheZ domain model, and domain queries for CheZ did not reveal new proteins missed by sequence searches. The CheZ sequences exclusively identified by sequence similarity searches contain the conserved catalytic residues, which supports their classification as CheZ proteins (Chapter 3.3.3).

2.3.2.4 FlaH and FlhA

The FlhA bacterial flagellum component shares similarity and a common origin with Type III secretion system proteins so convergence with PSI-BLAST analysis is not possible [116,117]. Similarly FlaH of the Archaeal flagella shares similarity to Type II secretion system components [118-120]. Despite these similarities the enzymatic role of these proteins gives them a high level of conservation, which makes them good for phylogenetic analyses of these two systems. To identify these components, BLAST queries with FlhA and FlaH were performed against the Bacterial and Archaeal genome sets of Microbial BLAST, respectively. Hits with scores of 256 or higher for FlhA and 300 or higher for FlaH were selected since the remaining sequences were paralogous to previous hits and were annotated as secretion proteins. Domain architecture analysis is not an effective tool for studying FlhA due to domain overlap with Type III secretion proteins, and there is no domain model for FlaH.

2.3.2.3 CheC, CheX, PilT, and PilU

The CheC and CheX proteins are members of a superfamily of phosphatases that also includes a flagellar protein FliY [27,121]. Members of this superfamily are distantly related to FliM. CheC, CheX, FliM, and FliY proteins were picked up in the CheC and CheX searches, but the latter two proteins were identified by C-terminal SpoA domains and removed from the final analysis. In addition, a split FliY protein from *T. maritima* that has a SpoA domain in a separate protein [121] was removed from analysis. The set of CheC and CheX sequences were initially aligned using ClustalX, and CheC and CheX proteins were further distinguished from each other based on secondary structure differences visualized by the VISSA program [110] and phylogenetic analysis that showed two distinct groups corresponding to the CheC and CheX proteins.

Similar to CheC and CheX, the PilU protein is a member of the superfamily of ATPases involved in Type II secretion systems and Tfp [122]. Similarity searches with PilU identify the PilU protein, its close homolog PilT [123-126], and more distantly related PilB and Type II secretion proteins [120,127]. PilT and PilU can be distinguished from the remaining proteins based on a multiple alignment of the entire set of sequences retrieved, because they lack a conserved metal binding insertion region with four conserved Cys residues [122]. Unlike CheC and CheX, the PilT and PilU proteins do not have obvious differences in secondary structure according to VISSA output. The two proteins were distinguished from each other by phylogenetic analysis of a multiple alignment of PilT and PilU proteins, which shows two distinct groups with overlapping phyletic distributions (Chapter 3.1).

2.3.2.4 CheY

CheY is a single domain protein encapsulating the ubiquitous REC domain [115]. Further complicating matters, stand alone REC proteins in some TCSs serve as middlemen in extended phosphotransfer relays [13,128]. In order to identify true CheY proteins, a BLAST query using the *E.coli* CheY protein was limited to retrieve sequences

between 100 to 150 a.a. in length to help restrict the results to single domain proteins. Although CheY is 129 a.a. in *E. coli* and 120 a.a. in *B. subtilis*, two divergent organisms, we extended the window approximately 20 a.a. in both directions to ensure the identification of the majority of CheY sequences. A single BLAST search with these parameters retrieved 1924 sequences. An exhaustive PSI-BLAST search was not performed under the assumption that the best BLAST hits would be the most closely related to CheY rather than other stand alone REC domains. Single domain REC proteins found within six orfs of other chemotaxis genes (*cheA, cheB, cheC, cheD, cheR, cheV, cheW, cheZ, or mcp*) were also retrieved to ensure that the length limit and single BLAST query did not exclude candidate CheY proteins. Once redundant sequences and sequences from redundant species and incomplete genomes were removed the remaining 999 sequences were aligned using MUSCLE [104] (Chapter 2.2.4). In the minimum evolution phylogenetic tree (Chapter 2.2.5) built from the final 886 set of single REC domain proteins (886 sequences), we identified two subfamilies with a large majority of members that are encoded in *cheA* gene neighborhoods. Sequences encoded near chemotaxis components that were not found in either of these subfamilies usually had another CheY candidate from the same gene neighborhood found within one of the two main chemotaxis associated subfamilies. All members of both subfamilies were retrieved for a set of 407 candidate CheY proteins, and an alignment of those revealed only seven proteins with extended C-terminal domains predicted to be involved in TCS outputs, which were removed resulting in a final set of 400 potential CheY proteins. A subfamily of these proteins have an undefined N-terminal domain, but they were kept in the set since they group tightly with other CheY proteins and are encoded near CheA. Although it is likely that this method still captured some stand alone REC proteins that are not CheYs, the overwhelming majority are assumed to be involved in chemotaxis. Since rigorous analysis of CheY was not the focus of this study, we chose to err on the side of over-inclusion.

2.3.2.5 CheA, CheV and CheW

CheW is a single domain protein (Pfam, SMART:CheW), whereas the multidomain CheA and CheV proteins contain CheW domains. CheA sequence searches using the full length protein did not pick up CheW or CheV proteins at high scores due to high conservation of its other domains. Domain architecture analysis of the resulting CheA set revealed that all CheA proteins contain the P1 hisitidine phosphotransfer domain (Pfam, SMART:HPT), the P4 ATPase domain (Pfam, SMART: HATPase_c) and the P5 CheW domain. Sometimes the P1 domain was found as a separate protein. Typically split CheA proteins are encoded adjacent to each other in their genomes, but when this was not the case, a blast query using the analogous region from the most closely related sequence (that is not split) based on phylogenetic analysis of the P3-P5 regions was used to identify the missing piece in the genome. Domain queries of the MiST database for all proteins containing CheW and HATPase_c domains revealed the same protein set identified by sequence searches.

CheV consists of a CheW and REC domain. Since REC domains are highly prevalent in two-component systems, a sequence query using the CheW protein was performed with a limit of 350 a.a. in length or smaller in order to pick up all CheV and CheW sequences while preventing hits with CheA proteins. A domain architecture query against the MiST database for all proteins that have CheW domains, but lack HATPase_c domains, revealed proteins with CheW domains that were not identified by the sequence search due to the length constraints. Some of these proteins contained multiple CheW domains, others contained N-terminal sensory domains identified by PSI-BLAST searches, and another is a divergent CheA protein containing a P1 and CheW (Chapter 3.2.1). Since only one protein with a CheW domain was exclusively identified by sequence searches (De.aro4 in Table A.10, a divergent CheW that was confirmed by gene neighborhood data), the final CheV set was retrieved by domain architecture queries for proteins with CheW and REC, but not HATPase_c, and the final CheW set was retrieved

by domain architecture queries for proteins with CheW, but not REC or HATPase_c (with De.aro4 included manually). The CheV query revealed one unusual protein with a N-terminal REC domain and a C-terminal CheW domain in *M. xanthus* (the opposite of the typical CheV domain architecture, Table 2.1) that was included in the CheW group instead. The final phylogenetic analysis of CheV and CheW revealed one CheV protein in *Bacillus anthracis* misidentified by our queries as a CheW protein due to a separation of the REC and CheW domains into two proteins.

2.3.2.6 MCP

The high conservation of the MCP signaling module is sufficient for MCP identification by domain query alone. MCPs from the final genome set (Table A.1) were identified by Pfam:MCPsignal and SMART:MA. The detailed MCP analysis was beyond the scope of this research project. A recent study by a member of our laboratory, Roger Alexander, revealed 12 length classes for MCP signaling domains [34]. He made HMMs for each length class from the multiple alignments built in his analysis. All MCPs for the final research set of this study were compared to the length class HMMs with HMMER by Roger. Sequences that matched a length class HMM with high scores and 2 gap spaces or less were classified as members of that length class. For the purposes of this study, it was not necessary to classify every MCP since the vast majority can be classified by the HMMs (Table A.13) even with these stringent guidelines. Only the 40H MCP class was used for further phylogenetic analysis for reasons specified in the results.

2.3.3 System Component Identification

Gene neighborhood, gene fusion, mirror tree, and phyletic distribution methods were used to identify associated proteins. In prokaryotes, gene neighborhood information is invaluable for inferring associated system components since interacting proteins and/or

members of a system are often encoded together [77]. Gene neighborhoods for each protein in this analysis are given in Tables A.2-17 and reveal that chemotaxis proteins are often encoded together. Gene fusions are also used to determine if two proteins interact [129-131]. We identified 80 CheA-CheY fusion proteins, 43 CheB-CheR fusion proteins, nine CheA-CheV fusion proteins, four CheW-CheR fusion proteins, two CheA-CheC fusion proteins, and one CheD-CheB protein in our sequence sets based on domain architecture analysis. Phylogenetic analyses of two interacting or associated proteins typically results in trees with similar topologies (mirror trees) due to co-evolutionary pressures [74,75]. Thus for systems where interactions are obscured by paralogs, similar subfamilies can be identified between the phylogenetic trees of chemotaxis proteins to aid specific component association. Mirror tree data is often complemented by gene neighborhood information, but even in the absence of such additional data, it is still a reliable indicator of system association. Similar to the mirror tree method, once subfamilies of chemotaxis systems were identified by analysis of CheA, CheB, CheR, CheD, CheV, CheW, CheY, and CheZ using the aforementioned methods, MCP length class information were associated with chemotaxis system subfamilies based on similarities in their phyletic distributions, a method known as phylogenetic profiling [132,133]. MCP class distribution was chosen over mirror tree methods in this case because the multiple gaps in MCP length classes result in very little sequence in common among the entire set, which heavily biases phylogenetic analysis when gaps are included, and does not provide enough sequence information when gaps are excluded. Since the 40H MCP class was associated with multiple chemotaxis subfamilies based on gene neighborhood data and phyletic distribution (Chapters 3.2.1 and 3.4), phylogenetic analysis of the 40H set was performed for mirror tree analysis of the class.

CHAPTER 3

UNDERSTANDING FUNCTION AND EVOLUTION OF THE CHEMOTAXIS SYSTEM THROUGH COMPARATIVE GENOMICS

Both flagella and type IV pili (Tfp) have been shown to confer motility in prokaryotes that possess either one or both. The bacterial flagellum is a complex organelle that consists of over 25 distinct proteins with multiple copies per flagellum and has sequence similarity to type III secretion systems [134]. The flagellum itself is a semi-rigid helical structure that is rotated to propel the bacterium forward. There is a great amount of diversity regarding the number and location of flagella in various organisms. The archaeal flagellum looks morphologically like the bacterial flagellum, but it is smaller and typically found as a tuft of polar flagella. Archaeal and bacterial flagellins and most other flagellar components are unrelated in sequence. Most of the components of archaeal flagella have not been clearly characterized, but the few that have bear a closer resemblance to Tfp than to bacterial flagellum components [135]. Tfp systems are related to type II secretion systems [120,134,136], but not all components of the Tfp system have been clearly characterized [137]. Tfp are typically bipolarly located, but only one pole at a time has an active motor as they work by extruding pili forward that attach to the surface and are retracted in order to pull the organism forward [136,138-140]. Unlike the bacterial and archaeal flagella, Tfp have also been shown to play roles in cellular competence, biofilm formation, and even electronic conduction [141-143].

Signal transduction systems regulate the majority of cellular processes in all organisms, and much like a car without a steering wheel, motility organelles make little sense to have without some way to control them. Class I histidine kinases (HKIs) regulate the majority of prokaryotic signal transduction through linear protein-protein

interactions as part of a TCS; however, class II histidine kinases (HKIIs) are a part of chemotaxis systems [144]. HKIs and HKIIs differ significantly in structure and in the proteins with which they interact. In TCSs, the kinase interacts with a response regulator made up of a receiver (REC) domain that is phosphorylated by the HKI and an output domain that can have a variety of functions, such as DNA binding for gene regulation or enzymatic activity for second messenger synthesis. The chemotaxis system kinase also phosphorylates a response regulator, but it is a stand alone REC domain that selectively interacts with the motility organelle based upon its phosphorylation state.

Decades of experimental work in *Escherichia coli* and *Bacillus subtilis* have classically portrayed the prokaryotic chemotaxis system as a protein interaction network that senses internal and external cues in order to regulate flagellar motility. Flagellar chemotaxis systems have been also begun to be characterized in many other organisms [48,57,61,145-151]. More recent studies in *Synechocystis* sp. PCC 6803, *Pseudomonas aeruginosa*, *Myxococcus xanthus*, and *Rhodospirillum centenum* have revealed chemotaxis systems that regulate motility via Tfp [50,123,140] and some that do not directly regulate motility, but rather have typical TCS outputs, such as regulation of gene expression [54], enzymatic activity [52], or other cellular functions [51,53]. A by product of functional diversity is that some organisms contain multiple chemotaxis systems that work independent of each other. Unexpectedly, experimental studies in *Rhodobacter sphaeroides*, *Synechocystis* sp., and *M. xanthus* have also revealed paralogous systems that affect the same motility [48,49,123]. The reason for and extent of such functional overlap is not clear and will undoubtedly continue to cause confusion in experimental studies of organisms with multiple chemotaxis systems.

Experimental work has not only revealed functional differences between chemotaxis systems, but also mechanistic differences in how the signals are transmitted and regulated through a varied repertoire of components [15,33]. The chemotaxis system of *E. coli* contains seven main components. It utilizes the core components CheA,

CheW, CheY, and MCP, in addition to the CheB/CheR adaptation system [152], and the CheZ phosphatase [153]. In *B. subtilis*, the chemotaxis system has all of the same components except for CheZ. An unrelated CheY phosphatase, CheC, is used in its place [121], and the CheC protein is activated by interactiing with CheD, an enzyme that deamidates specific sites on the MCPs making them accessible for methylation by CheR [29,36]. In addition to CheW, *B. subtilis* has an additional scaffolding protein, CheV [154], which consists of CheW and an additional REC domain that is phosphorylated by CheA [37]. Although not present in either of the aforementioned model systems, the third CheY phosphatase, CheX, has been studied experimentally in *Treponema denticola* [149], *Borrelia burgdorferi* [155], and *Thermotoga maritima* [27]. CheX is homologous to CheC in both sequence and structure, but it acts as a dimer and does not interact with CheD [27]. While many of these proteins have been studied in flagellar systems, the component repertoire of systems that regulate Tfp motility or alternative outputs have proven thus far to be much simpler based on experimental and genomic analysis [47,52,54,156]. Given the wealth of available data at the experimental and genomic levels, this system is ideally suited for in-depth computational genomics study to elucidate the scope of the system's diversity as well as its evolutionary history.

3.1 Three Functional Families of the Chemotaxis System

The multidomain CheA kinase is a highly conserved essential component of the chemotaxis system with only one member per system making it an ideal subject for phylogenetic analysis. Although classically defined as a five domain protein (P1-5), we have found extensive domain architecture variability in CheA consisting of duplications, deletions, and fusions surrounding a conserved core (P3-P5) [22]. The P3 domain is responsible for dimerization [22,157], the P4 domain is a conserved ATP-binding domain [22,158], and the P5 domain is homologous to the CheW protein [24] and is involved in scaffolding interactions with MCPs and CheW proteins [159]. In addition to these three

domains, the P1 domain, a conserved phosphotransfer domain, is also essential to all CheA proteins because it is the site of CheA autophosphorylation prior to phosphotransfer to CheY and CheB [32,160]. The P2 domain binds CheY to facilitate phosphotransfer to the P1 domain [20,21,26]. A comparison of the two-dimensional domain architecture with the three-dimensional structure shows the P3-P5 domains correspond to a distinct globular unit [22] and the current dimerization and Hpt domain models (Pfam:H-kinase_dim and Hpt, respectively) do not capture the entire length of their associated domains (Figure 3.1). The H-kinase_dim domain model is not built from the full dimerization region; whereas the Hpt domain fails to capture the fifth helix of the

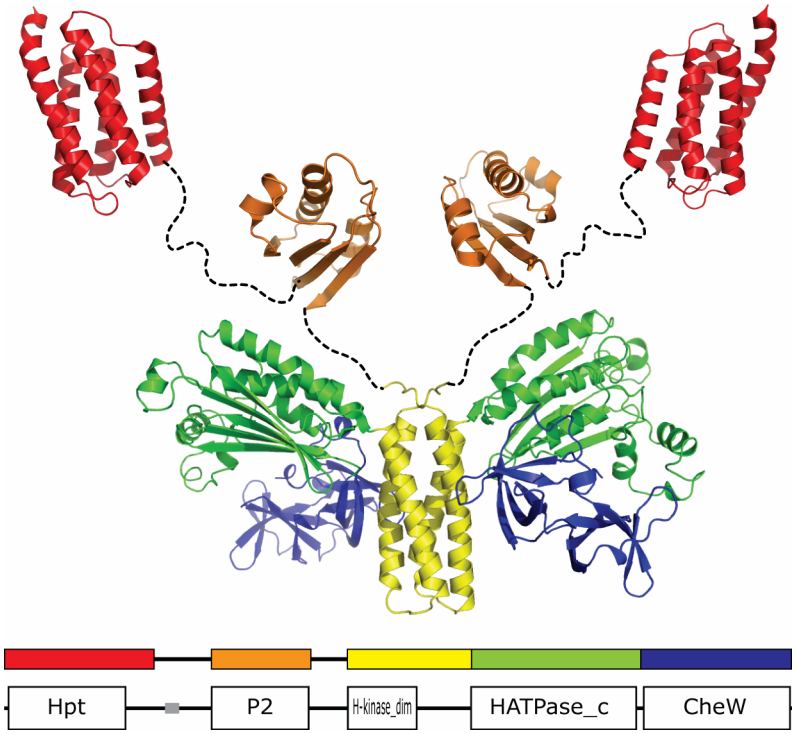


Figure 3.1 The CheA domain architecture captures information about its three-dimensional structure. The Pfam domain model of CheA (GI 15643465) visualized in the MiST database [87] and its two-dimensional color scheme are shown below the three-dimensional model that has a matching color code. The three-dimensional model consists of three different crystal structures: the P1 in red (PDB, 1I5N), the P2 domain in orange (PDB, 1UOS), and the P3-P5 domains in yellow, green, and blue, respectively (PDB, 1BDJ), with the linker regions manually added as dashed lines.

P1 domain identified in the crystal structure, which is absent in the Hpt domains originally crystallized from TCSs [32].

The multidomain nature of CheA and the tendency of CheA proteins to be encoded adjacent to other chemotaxis components on the chromosome [48,49,51,53-65] provide several independent methods for the validation of a CheA phylogenetic tree. The CheA maximum likelihood tree built from the conserved P3-P5 core shows three distinct groups. Further analysis based on gene neighborhood, experimental data, and the presence or absence of Tfp and flagella supports that these groups are formed due to differences in their outputs. One family regulates flagellar motility whereas members of the other two families regulate Tfp motility and other TCS-like outputs (alternative outputs), respectively (Figure 3.2).

Genome analysis showed that 116 of the 122 organisms that contain flagellar-type CheAs also possess flagella as defined by full-length FlhA or FlaH proteins encoded in their genomes, which are conserved components of the bacterial and archaeal flagellum, respectively. Mutations in the flagellar CheAs and/or their neighboring components in *E. coli* [161], *B. subtilis* [162], *R. sphaeroides* [40,48,163], *Azospirillum brasilense* [61], *Sinorhizobium meliloti* [164], *Halobacteria salinarum* [165], *Vibrio cholerae* [57], *Listeria monocytogenes* [148,166], *T. denticola* [167], *Ralstonia solanacearum* [150], *Helicobacter pylori* [145], and *Caulobacter crescentus* [168], *P. aeruginosa* [62], *Pseudomonas fluorescens* [169], and *R. centenum* [51,60] have resulted primarily in flagellar motility defects (Figure 3.2). Six organisms that do not have flagella (as defined by the absence of FlhA in the proteome) possess flagellar-type CheA. The absence of flagella in two *Shigella* species and two *Bordetella* species is predicted to be the result of recent system loss as adaptations to their hosts, and in accordance their chemotaxis systems are likely to be lost in the future. *Erythrobacter litoralis* lacks FlhA and has a highly divergent CheA (Er.lit in Figure 3.2) in comparison to its closest relatives. We

predict that this system is in the process of being removed from the genome. Non-flagellated *M. xanthus* surprisingly has two flagellar chemotaxis systems (My.xan1 and My.xan8 in Figure 3.2); however, this species is the only member of sequenced δ-proteobacteria thus far to lack FlhA, including the very closely related *Anaeromyxobacter dehalogenans*. Both flagellar systems in *M. xanthus* group with those from other δ-Proteobacteria, which supports the notion that the two flagellar chemotaxis systems may have evolved new functions after the loss of the flagellar proteins. 28 of the 32 organisms that have Tfp CheAs also have Tfp (as defined by a PilU protein encoded in their genomes), and studies in *P. aeruginosa* [50] and *Synechocystis* sp. [63] have shown three members to affect Tfp related motility. In the three *Pseudomonas* species that have recently lost PilU, we predict that their Tfp chemotaxis systems will follow the same fate or take on a new function. Genomic distribution of Alt systems do not show a correlation with either Tfp or flagella, which is consistent with their association with outputs other than direct motility regulation (My.xan5 [54], Rh.cen3 [53], and Ps.aer4 [52] in Figure 3.2). The observation that all of the organisms with Alt chemotaxis systems possess flagella and/or Tfp suggests that the alternative output family arose later in evolution and Alt systems might still play indirect roles in motility regulation.

Currently all of the flagellar CheAs that have not been experimentally shown to play a role in chemotaxis are found in organisms that have multiple chemotaxis systems thus being classic examples of non-orthologous gene displacement [170]. As evidenced by studies in *R. centenum*, functional paralogs can diverge resulting in alternative outputs (see Rh.cen1 [60] and Rh.cen2 [51] in Figure 3.2). The laterally transferred flagellar system of *M. xanthus* (My.xan8 in Figure 3.2) has been experimentally characterized as the DiF (Defective in Fruiting body) system [59]. Although the DiF system affects Tfp motility in *M. xanthus* and possibly in the other δ-Proteobacteria that have this system, it has a flagellar-type origin that is explained further in Chapter 3.4.1. Experimental studies in Dif show that it regulates fibril biogenesis [171], which supports the hypothesis that its

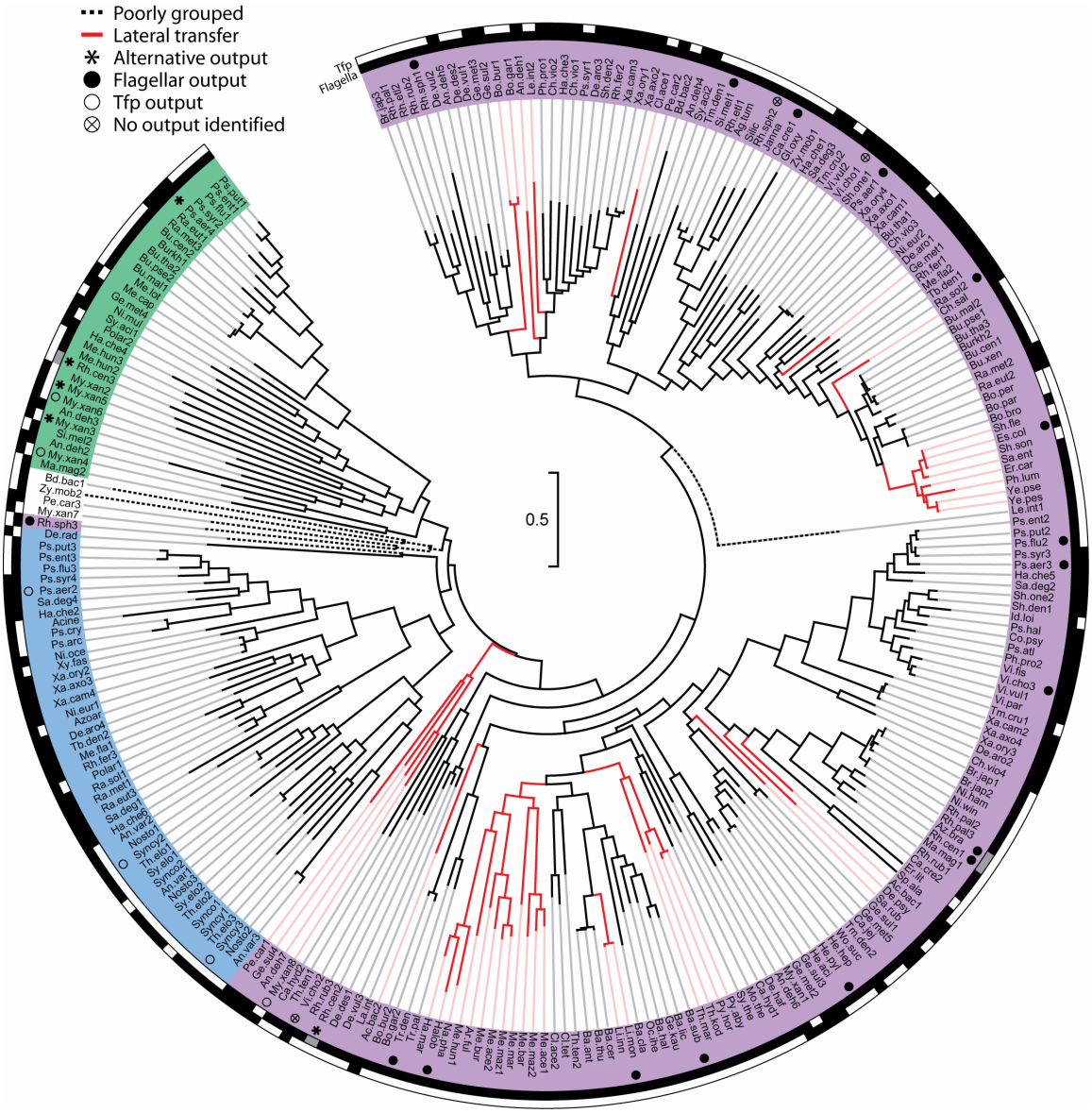


Figure 3.2 Phylogenetic analysis of CheA reveals three distinct functional families supported by experimental data and coorelations with Tfp and flagella. The three families regulate flagellar motility (purple), Tfp motility (blue), and alternative outputs (Alt) typically associated with TCSs (green). Available experimental data for a sequence is shown next to its identifier, and black markings in two rings around the tree shown sequences that come from organisms with Tfp (outer ring) and/or flagella (inner ring). Lateral transfer is not shown in the Alt chemotaxis family because it is readily laterally transferred, which obscures its evolutionary history unlike the flagellar and Tfp systems. The CheA of *Halobacterium* sp. NRC-1 differs only by one amino acid from the experimentally characterized CheA from *H. salinarum* [165], and the two proteins are assumed to be functionally equivalent. Identifiers correspond to Table A.2.

effect on Tfp based motility is indirect. *M. xanthus* is also unusual in that it has Alt systems that affect Tfp motility [49,140]; however, these results have not been proven to be due to direct motility regulation, rather than indirect consequences of alternative outputs.

In keeping with the tight correlation between flagella and flagellar chemotaxis systems, 103 out of 116 organisms that contain FlhA and 13 out of 21 organisms that contain FlaH sequences also possess at least one flagellar-type chemotaxis system. The flagella of *Legionella pneumophila* have been shown to be involved in virulence [172], which supports its maintenance despite a lack of any chemotaxis system. Similarly, the flagellum of a free-living bacterium *Hyphomonas neptunium* is functional [173] and appears to aid the dispersal of cell populations without a dedicated chemotaxis system. The remnants of this system in *H. neptunium* (CheR and CheY only) suggests its recent loss. Endosymbionts such as *Buchnera aphidicola*, *Wigglesworthia glossinidia*, and *Sodalis glossinidius* have lost most of their chemotaxis components (including CheA) and are predicted to be in the process of losing their flagellar components due to rapid sequence evolution from their small effective population sizes [174]. The stronger correlation between the presence of bacterial flagella and associated chemotaxis systems in comparison to those of archaea is likely due to the foreign origin of archaeal chemotaxis systems. Archaea received chemotaxis systems by lateral transfer like their TCSs [175]. Taking into account the fact that bacterial and archaeal flagellar-type chemotaxis systems are homologous and closely related, whereas their flagellar systems are unrelated, it is plausible that both archaeal and bacterial flagella are evolutionarily older than the chemotaxis system and have functions beyond motility.

PilU and its homolog PilT have been shown to play roles in the retraction of Tfp, but PilU has been implicated to play more of a regulatory role in Tfp systems of *Neisseria gonorrhoeae [124]*, *P. aeruginosa [126,139,176]*, and *Synechocystis* sp. [123], which leads us to choose PilU as a marker for Tfp that are capable of regulation by

chemotaxis components. Genome analysis of PilT and PilU distribution revealed that 60 of the 62 organisms that have PilU also have PilT. 33 out of the 93 organisms that have PilT lack PilU, including organisms that have type IVB pili associated with pathogenicity [177]. We predict that Tfp systems that exclusively utilize PilT are involved in processes that require less regulation that those with PilU or are in the process of being lost. Further complicating matters, Tfp are known to be involved in processes other than motility [141-143], which makes it impossible to establish a 1:1 relationship between Tfp motility and chemotaxis systems. Only 28 of the 61 organisms that have PilU also have Tfp chemotaxis systems, which shows that even though PilU plays a regulatory role it might not be through direct interactions with chemotaxis components. Regardless, flagellar and Tfp chemotaxis systems have predominantly been propagated by vertical inheritance in accordance with the evolutionary history of both motility systems as seen by their locations on a 16S rRNA phylogenetic tree (Figure 3.3). The evolutionary history of the Alt systems is obscured by lateral transfer events since is not under pressure to co-evolve with a motility organelle, much like TCSs [178].

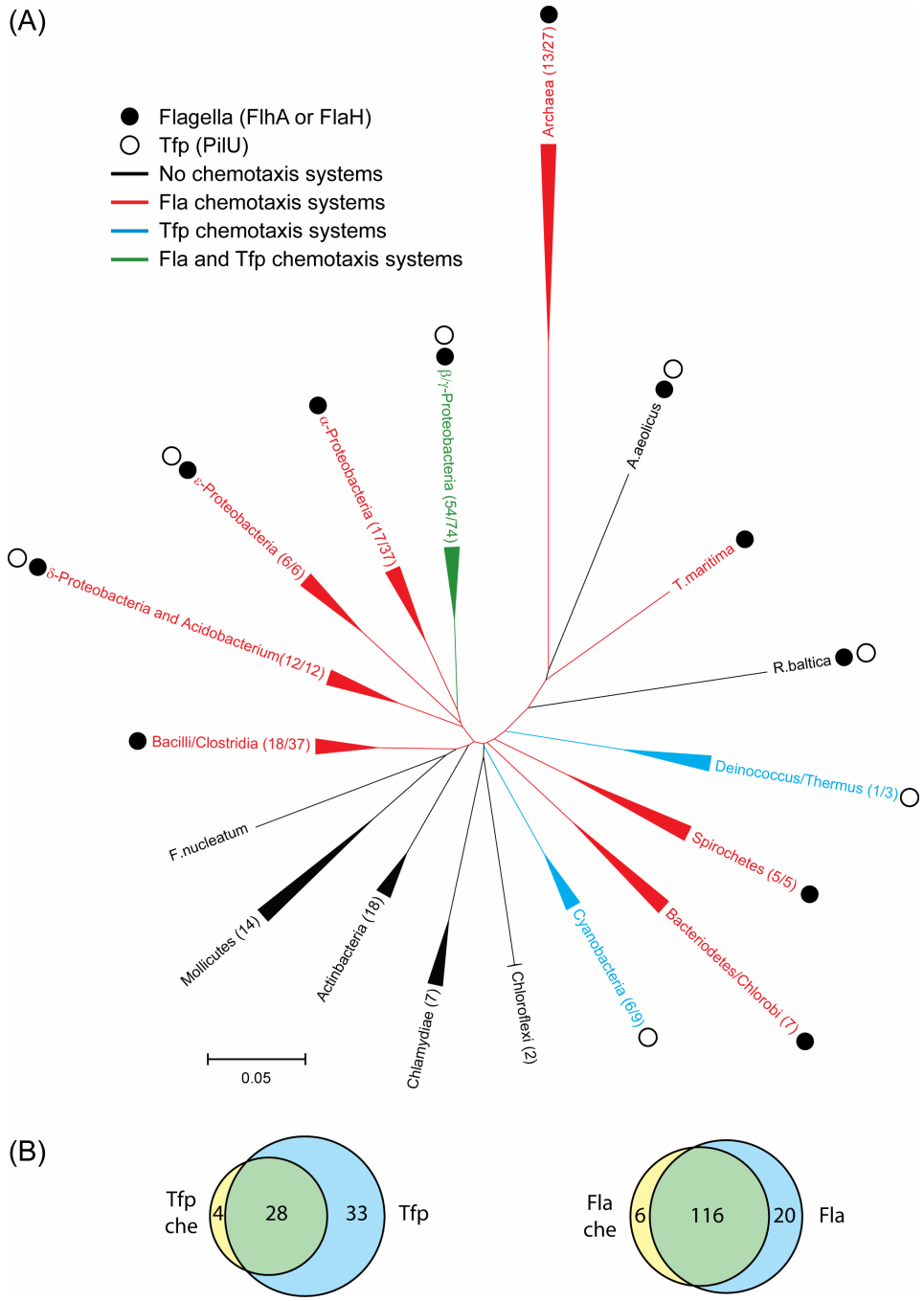


Figure 3.3 (A) The phyletic distribution of Tfp and flagellar chemotaxis systems and motility organelles shows a pattern of co-evolution on a 16S rRNA tree. Clades where flagellar, Tfp, or both chemotaxis systems are present are shown in red, blue, and green, respectively. The number of species in each collapsed clade is given, and for clades with chemotaxis systems, the number of chemotaxis systems is given as a ratio of the number of species. Organisms correspond to Table A.1. **(B)** Venn diagrams (not to scale) showing the partial overlap (green) of the numbers of organisms with Tfp chemotaxis systems (yellow) and those with Tfp motility systems (blue) and the partial overlap (green) of organisms with flagellar chemotaxis systems (yellow) and flagellar motility systems (blue).

3.2 Chemotaxis Family Characterizations

3.2.1 CheA Domain Architecture Diversity

All CheA proteins identified in our analysis have identifiable Hpt, dimerization, HATPase_c, and CheW domains, but there is extensive domain architecture diversity within this protein family. Only Alt-type CheAs (CheA-Alt) show highly conserved domain architecture reinforcing the notion that they are a recent addition to the CheA family. Although CheA is classically described as the five domain protein shown in Figure 3.1, CheA-Alt proteins lack the P2 domain. The dimerization domain of CheA-Alt is slightly longer than that of flagellar-type CheAs (CheA-Fla) and poorly identified by the current dimerization domain model (Pfam:H-kinase_dim). They also have a C-terminal REC domain; a configuration that has resulted in such hybrid proteins being described as CheAY proteins. Similar to CheA-Alt, Tfp-type CheAs (CheA-Tfp) have a hybrid (C-terminal REC) domain architecture and lack CheA. All CheA-Tfp proteins have an elongated dimerization domain with a conserved sequence motif (Figure 3.4). The conserved motif occurs in the middle of the CheA-Tfp dimerization domain corresponding to the region between the two α-helices of the domain (Figure 3.1), making it available for possible interactions with other proteins that have not yet been identified. Despite these similarities, the Tfp CheAs from cyanobacteria show different domain architectures compared to those of β/γ-Proteobacteria.

Figure 3.4 Sequence logo [112,179] of the CheA dimerization motif.

An experimentally described CheA-Tfp protein in *P. aeruginosa* (Ps.aer2 in Figure 3.2) was found to have eight Hpt-like domains [50]. Two of the eight Hpt-like domains lack the conserved histidine residue used in the phosphotransfer reaction of the domain, and instead have serine or threonine. In addition to the divergent Hpt domains, the N-terminal region of Ps.aer2 was described as having a region with homology to FimL [50], a protein involved in Tfp biogenesis {Whitchurch, 2005 #1047. All of the CheA-Tfp proteins from β/γ-Proteobacteria have multiple Hpt domains and an N-terminal region homologous to FimL (with the exception of three divergent Tfp chemotaxis systems that have truncated CheAs, Acine, Ha.che6, and Sa.deg1 in Figure 3.2). A closer analysis showed divergent Hpt domains in all of the full length CheA-Tfps from β/γ-Proteobacteria based on PSI-BLAST searches of undefined regions and secondary structure prediction.

It was predicted that the two divergent Hpt domains of Ps.aer2 are still phosphorylated at their alternative residues, but sequence analysis of the divergent Hpt domains of our data set does not show conservation of the serine and threonine active site residues. We predict that the divergent Hpt domains are involved in binding CheY in a manner analogous to the P2 domain described in other CheA proteins. PSI-BLAST search with the FimL-like region, excluding the flanking Hpt and Hpt-like domains, confirmed its similarity to the FimL, and also suggested that there are two domains within the region. One shows homology to the previously identified divergent Hpt domains of the CheA-Tfp subfamily, and the other shows homology only to similar regions in other Tfp CheAs and to a distinct region in FimL. Secondary structure analysis of the unique region predicts five α-helices. Given the extensive duplication and divergence of the Hpt domain in this protein family, the unique region is consistent with an extremely divergent Hpt domain.

The cyanobacterial Tfp CheAs have only one Hpt domain; however, closer analysis revealed all of these proteins have an additional domain following the N-terminal Hpt domain. The domain showed no homology to other domains in PSI-BLAST analysis, but secondary structure showed that it consists of five α-helices. Given the secondary structure and the relationship between these CheA proteins and the Tfp CheAs of β/γ-Proteobacteria, we predict that these are divergent Hpt domains that bind CheY. We have termed the divergent Hpt domains found in all Tfp CheA proteins to be P2-Hpt domains, to reflect their predicted function and origin. Beyond the P2-Hpt domains, cyanobacterial Tfp CheAs have an undefined region of variable length that contains repeats in some sequences, but lacks domains common among all of them. There is a frameshift in the *Deinococcus radiodurans* Tfp CheA, and its Hpt domains are encoded in an unannotated ORF before the rest of the gene. The translated product of the ORF consists of two Hpt domains, which is consistent with the presence of multiple Hpt related domains in all CheA-Tfp proteins. Although there can be a variable number of Hpt and P2-Hpt domains, we have found that all Tfp CheA from β/γ-Proteobacteria (except for the three divergent sequences mentioned previously) contain four Hpt domains, one at the N-terminal included in the FimL-like region and three following the FimL-like region. Additional Hpt and Hpt-like domains can be inserted between the three last Hpt domains and before the dimerization domain. All of these CheA sequences also contain the three P2-Hpt domains associated with the FimL region, except for five sequences that have lost the most divergent domain.

Flagellar CheA proteins show a range of unrelated domain architecture. The experimentally characterized flagellar CheA protein of *H. pylori* is a CheAY fusion {Jimenez-Pearson, 2005 #335} that lacks a P2 domain like CheA-Alt. A flagellar CheAY protein was also described in *R. centenum* (Rh.cen1 [60] in Figure 3.2), but a closer analysis reveals that it is instead a CheA-CheV fusion protein. The second CheW domain is divergent and was not recognized in that study. Although a CheA-CheV

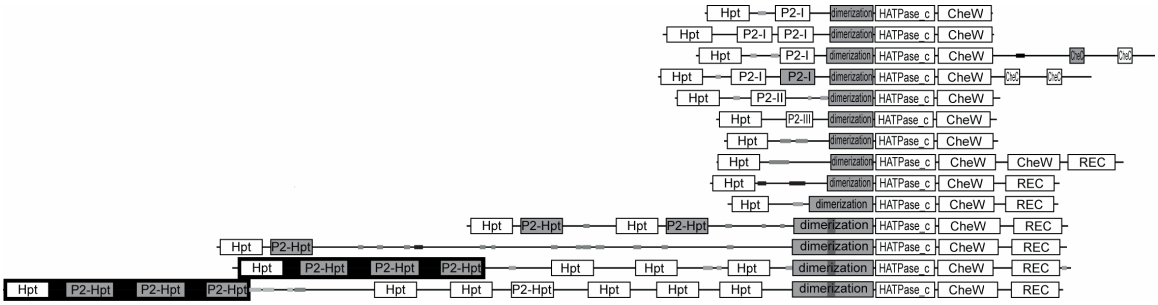


Figure 3.5 A common core and diversity of CheA homologs. The domain architectures of the flagellar CheAs Ba.sub, Py.aby, Le.int1, Me.hun1, Si.mel1, Es.col, Ps.aer1, Br.jap1, and He.pyl; the Alt CheA Ps.aer4; and the Tfp CheAs Syncy3, Nosto1, Ra.eut3, and Ps.aer2 are shown from top to bottom. Sequence identifiers correspond to Figure 3.2 and Table A.2. All shown domains are from Pfam as visualized in MiST, but the REC moniker from SMART is used in place of Response_reg from Pfam for legibility. All domains in gray were identified by manual sequence analysis. Although the dimerization domain was partially identified in some sequences, the HMM does not cover the length of the domain model. The P2-Hpt domain shown in white was identified as an Hpt domain by the Pfam domain model, but it lacks the conserved histidine. Regions homologous to the FimL protein are surrounded by a black box. Small black, gray, and white boxes indicate predicted coiled-coil, low complexity, and signal peptide regions, respectively. Despite diverse domain architectures, all CheA proteins contain Hpt, dimerization, HATPase_c, and CheW domains, with the latter three forming a tight core consistent with its structure shown in Figure 3.1.

domain configuration could also be achieved by CheW duplication and CheY fusion, the analysis of the evolution of the whole system discussed in Chapter 3.5 supports the hypothesis that this domain architecture is the result of a CheA-CheV fusion event. Domain architecture analysis of flagellar CheAs also revealed the existence of CheA-CheC fusion proteins, and some CheA proteins that contain only Hpt, dimerization, ATPase, and CheW domains (Figure 3.5).

Crystal structures of P2 domains from *E. coli* and *T. maritima* show significant differences between the two [26]. Sequence analysis of the P2 domain region further revealed three classes of related P2 domains (Figure 3.6). P2 domains are exclusively found in CheAs from flagellar systems, and in some of these CheA proteins the P2 domain has been duplicated (Figure 3.5). One subfamily of CheA proteins has a conserved repeat in the P2 region that is predicted to bind CheY based on its location

since the majority of CheA proteins are predicted to have a P2 or P2-Hpt domain. An unusual CheA was experimentally identified in *R. sphaeroides* that was characterized as having only an Hpt domain and a CheW domain tethered by a long linker region [163]. The unusual CheA works with a truncated CheA in the same gene neighborhood that lacks an Hpt domain and a P2 domain. Sequence analysis of the Hpt-CheW CheA revealed a P2-II domain in the linker region. Diversity of the CheA domain architectures is shown in Figure 3.5.

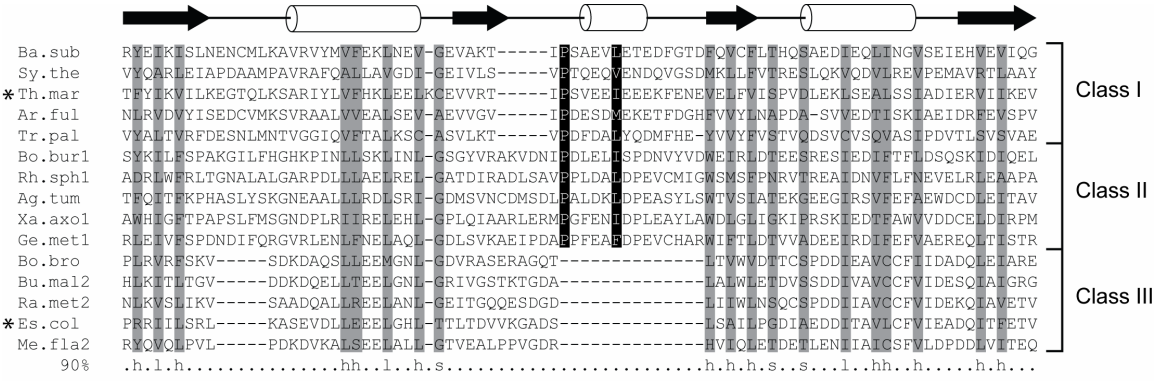


Figure 3.6 Multiple alignment of the P2 domain and its classification. Three subclasses of the P2 domain were identified. A multiple alignment with representative members of each class of P2 domain shows the insertions and deletions that define each class. Positions conserved at 90% or more (excluding turn-like residues) in an alignment of 129 P2 sequences are shown in gray. Conservation consensus is shown underneath the alignment (h, hydrophobic; l, aliphatic; p, polar; s, small). Black columns show conserved proline and hydrophobic positions in classes I and II. The secondary structure elements are shown above the alignment based on crystal structures from *E. coli* and *T. maritima* [20,26] and the seqeunces associated with the structures are identified by asterisks. Black arrows represent β strands. White cylinders represent α helices. Sequence identifiers correspond to those found in Table A.2.

3.2.2 Characterization of CheA Functional Classes

3.2.2.1 Tfp and Alt systems

The Tfp systems show a highly conserved gene order that typically consists of two *cheY* followed by *cheW*, *mcp*, and *cheA* genes (Figure 3.7), and has a distribution in *D. radiodurans*, Cyanobacteria, and β/γ-Proteobacteria. Some divergent CheB and CheR proteins are found to be encoded in the gene neighborhoods of a few Tfp systems from γ-Proteobacteria, but have not been shown to be essential for function when deleted in *P. aeruginosa* [50,180]. Although a Tfp system is present in *D. radiodurans*, it cannot be clearly established whether it originated there or was gained by lateral transfer given the organism's propensity for DNA uptake [181]. The closely related *Deinococcus geothermalis* and *Thermus thermophilus* also have Tfp (Figure 3.3), but both lack chemotaxis systems. The CheA-Tfp of *D. radiodurans* groups separately from the those of cyanobacteria and β/γ-Proteobacteria (Figure 3.2), which supports that it was vertically inherited.

The Alt systems typically encode two CheW proteins, one MCP, CheA, a TCS-like response regulator that consists of a REC domain fused to an output domain (unlike CheY), CheB, and CheR protein with tetratricopeptide repeat (TPR) domains fused to its C-terminus, which are not found in other types of CheR. TPRs mediate a wide variety of protein-protein interactions [182]. The gene order is conserved to a limited degree in most Alt systems, often featuring *cheW* followed by *cheR* and the second *cheW* and *cheB* following *cheA* with *mcp* before or after the WRW group (Figure 3.7). There is not consistent taxonomic grouping within the Alt group, which is typical of proteins that are subject to frequent lateral transfer events. The propensity of this system for lateral transfer is predicted to be linked to its TCS-like outputs because this system is not constrained by interaction with a motility organelle. Five CheA sequences group with the Tfp and Alt systems on the tree,

Figure 3.7 Topology only representation of the CheA tree from Figure 3.4 that shows the gene neighborhoods for each sequence and flagellar subfamily groupings. Branches in red indicated laterally transferred flagellar-type Ches. Dashed branches indicate poorly resolved sequences. Although some flagellar subfamily delineations are not entirely consistent with gene neighborhood, such as that between 7a and 7b or the large inclusion of members in the flagellar 1 family (F1), they are consistent with additional data such as domain architecture and phyletic distribution. Going clockwise from the top the gene neighborhoods of F8 through the largest F1 group are in order from the inner circle to the outer circle, and F2-Alt are in order from the outermost to the innermost. A few gene neighborhoods have been truncated for legibility. Detailed information about the class and gene neighborhood of each identifier shown in Figure 3.2 can be found in Table A.2.

but they do not possess conserved features of either system. One of the five systems has been shown to control flagellar motility in *R. sphaeroides* (Rh.sph3 [48,163] in Figure 3.2). These highly divergent systems are predicted to be rapidly evolving as indicated by their long branch lengths and are thus far unique, resulting in their inability to be clearly resolved on the tree.

Neither the Tfp nor Alt systems show any correlation with CheC, CheD, CheV, CheX, or CheZ proteins based upon gene neighborhood analysis. A recent study has revealed 12 classes of MCPs defined by differences in the length of their signaling domains [34]. Analysis of the MCPs in the gene neighborhoods of Tfp and alternative output (Alt) systems and those in organisms that only have Tfp and/or Alt systems shows that only the 40 heptad (40H) MCP length class is clearly associated with Tfp and Alt systems. The 40H MCP class is associated with multiple flagellar subfamilies as well as with the Tfp and Alt families. The three functional families can be clearly distinguished on a phylogenetic tree of the 40H MCP class that also confirms the close relationship of the Tfp and Alt systems shown in the topology of the CheA tree. Although it cannot be established if the Tfp chemotaxis system of *D. radiodurans* was inherited vertically or horizontally, the phylogenetic analysis of its three MCPs (all 40H) puts them in a monophyletic cluster, which supports a scenario of vertical inheritance with subsequent duplication given the predominance of a single MCPs associated with these systems. Of the 35 MCPs found in organisms that lack any flagellar CheAs, 31 are classified as 40H. Furthermore all of the 44 MCPs in the Tfp group are encoded in gene neighborhoods with Tfp *cheAs*. 27 of the 29 MCPs in the Alt group are encoded in gene neighborhoods with Alt *cheAs,* and all 29 Alt MCPs are in organisms with Alt CheAs encoded in their genomes (Figure 3.8). Unlike flagellar systems, the majority of Tfp and Alt systems have only one MCP associated with each system/CheA.

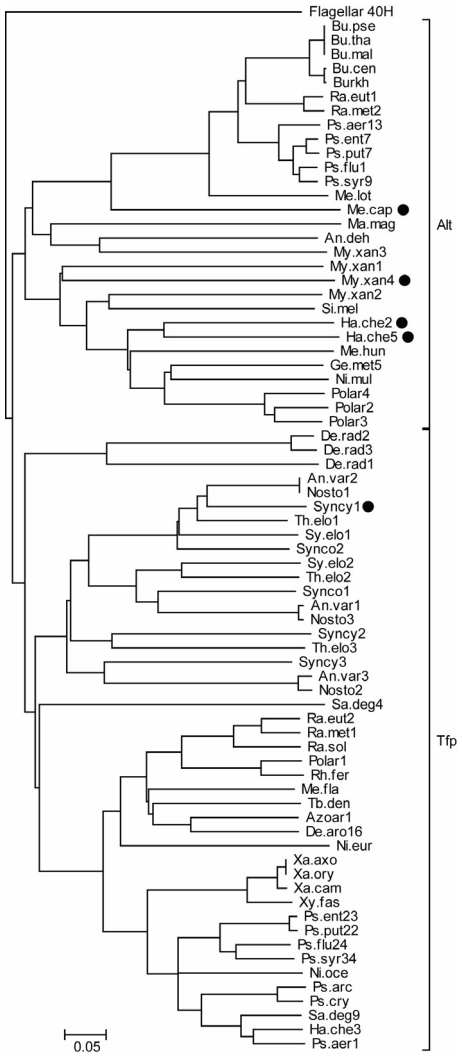


Figure 3.8 Phylogenetic analysis of the 40H MCP class shows distinct groups associated with the Tfp and Alt chemotaxis families based on gene neighborhood data. As seen in the CheA tree (Figure 3.2) Tfp and Alt MCPs group together. Black circles respresent sequences not encoded in a CheA gene neighborhood, but all members of the Tfp and Alt groups are from organisms that encode their respective systems. Sequence identifiers correspond to Table A.12.

3.2.2.2 Flagellar systems

As shown in Figures 3.2 and 3.7, the flagellar family of chemotaxis systems is much larger than even the Tfp and Alt systems combined, thus far. The flagellar chemotaxis family has extensive diversity as demonstrated by variable gene neighborhoods (Figure 3.7) and accessory components. The strong co-evolutionary

pressure on flagellar chemotaxis systems and their associated organelles has resulted in a pattern of vertical inheritance with discrete instances of lateral transfer (Figure 3.2). We have identified 10 distinct flagellar classes in the CheA tree that are corroborated by gene neighborhood similarities (Figure 3.7), phyletic distribution, correlations to MCP length classes (Chapter 3.4), and CheA domain architecture. Although not all members of flagellar class 1 (F1) or 4 (F4) group together on the CheA tree, their associations are supported by analysis of the remaining chemotaxis components that are explored in the following section (Chapter 3.3). The uncategorized system of *Salinibacter ruber* has some characteristics similar to the F3 and F4 systems it groups with on the CheA tree, but it has also diverged significantly and cannot be categorized with either. Similar to the ununsual *R. sphaeroides* system (Rh.sph3 in Figure 3.2), an unusual CheA in *Leptospira interrogans* (Le.int1 in Figure 3.2) does not clearly associate with any group, but analysis of other chemotaxis proteins allows us to classify it as a member of the flagellar 1 (F1) family. The numbering of the subfamilies is delineated in Chapter 3.4, and the further subclassifications of some are based upon chemotaxis component repertoire differences that will be covered in the next section.

3.3 Component Analysis

3.3.1 CheB and CheR Analysis

The CheB methylesterase and CheR methyltransferase work together to regulate the methylation state of MCPs. CheB typically consists of two domains: an N-terminal REC domain is phosphorylated by CheA to regulate the activity of the C-terminal catalytic domain. Domain architecture analysis of CheB revealed multiple sequences that lack the REC domain and confirmed previously identified large fusion proteins that consist of the CheB catalytic domain (Pfam:CheB_methylest), CheR, multiple PAS domains, and often domains associated with HKIs [183,184]. In addition to fusions with

CheB catalytic domain, CheR has also been found to be fused to TPR and CheW domains [184]. An unusual CheD-CheB fusion is found in *Bdellovibrio bacteriovorus*. Phylogenetic trees built from the conserved catalytic core of CheR and CheB domains showed roughly similar topology to the CheA tree, but the resolution was poorer due to the short length in comparison to the multidomain core of CheA. Since 250 of the 280 CheB proteins and 292 CheR proteins can be clearly associated with their partners using gene neighborhood, gene fusion, and mirror tree methods, a tree was built from a concatenated alignment of the core catalytic regions of the 250 pairs to increase the tree resolution. Of the 250 pairs, 220 are in the same gene neighborhood or are fused together. Only the F1 and F6 CheB and CheR pairs are found separate from each other, but the tight phylogenetic clustering of each group made it easy to resolve the pairs using the mirror tree method. Like the CheA tree, the CheBR tree shows Tfp, Alt, and flagellar functional groups, but it also possesses two groups that are associated with TCSs rather than chemotaxis systems. One TCS family is made up of the CheB-CheR fusion proteins that are often also fused to and/or encoded adjacent to HKI catalytic modules. The members of the other TCS family are almost always encoded adjacent to HKIs. All TCS associated CheB components lack the N-terminal REC domain typically found in CheB proteins, and some TCS subfamily members are found in organisms that lack other chemotaxis components. A few of the fusion proteins have an extra CheB protein (catalytic domain only) encoded adjacent to them, which has a C-terminal extension with a conserved motif. The extra CheB proteins lack a cognate CheR and are assumed to work with the fusion proteins.

The revelation of CheB and CheR proteins that are associated with TCSs instead of chemotaxis systems begs the question of the functional relationship between the proteins. TCSs do not have any components that are analogous to the MCPs so we wanted to determine if these chemotaxis components are working directly with the associated TCSs. The CheB-CheR fusion proteins have a conserved N-terminal

consisting of the CheB catalytic domain, CheR, a divergent PAS domain, a long coiled-coil region, and another PAS domain. The remaining portion of the protein is variable often with more PAS domains, some with HKI signalling modules, and others with DNA-binding domains or other outputs. PAS domains are involved in sensing multiple stimuli, often oxygen concentrations and redox potential, in order to regulate many important cellular processes [185]. One of the fusion proteins has been experimentally studied and found to have a previously undefined histidine kinase module due to large amounts of sequence divergence [183]. The HKIs in the neighborhood of the other family of TCS associated CheB and CheR proteins have a variable N-terminus that usually consists of a periplasmic CHASE3 sensing domain [186]or a stretch of multiple HAMP domains [187] followed by the conserved core made up of a GAF domain [188], a coiled-coil region, the catalytic modules, and three REC domains.

In chemotaxis systems, CheB and CheR act on conserved methylation sites [34,189] of the MCP signaling module, a four helix coiled-coil [23,31]. Further analysis of the TCS CheB and CheR associated kinases shows that the coiled-coil regions within them often have multiple repeats of a motif that is very similar to the chemotaxis methylation site motif. The MCP methylation motif is [ASTG]-[ASTG]-[X]-[EQ]-[EQ]-[X]-[ASTG]-[ASTG] [34], whereas the methylation-like sites in the CheB-CheR fusion coiled-coils have a consensus of [AST]-[X]-[X]-[E]-[E]-[X]-[X]-[AST]. The coiled-coils of the gene neighborhood associated HKIs have a looser consensus of [ATGQV]-[X]-[X]-[E]-[E]-[X]-[X]-[ATGQV]. The two absent small positions in the HKI coiled-coils correspond to positions in the MCPs that are important for packing the four-helix bundle rather than for direct CheB or CheR interaction. Unlike MCP signaling domains, the TCS coiled-coils are predicted to be two-helix bundles, consistent with the N-terminal coiled-coil region of the crystalized catalytic module of an HKI [190]; thus, the small position is not conserved. We propose that the methylation of these residues may affect the rigidity and conformation of the coiled-coil and thus alter the transmission of signals

between the domains it links. Experimental work is needed to validate the predicted interaction and function of these proteins, but the information thus far does support a direct interaction between the divergent CheB and CheR proteins for an output unrelated to chemotaxis.

The few CheB proteins that are associated with Tfp systems have divergent REC domains that lack the conserved aspartate site of phosphorylation. It has been shown that knocking out the Tfp CheB and CheR proteins in *P. aeruginosa* does not affect Tfp based motility, unlike knockouts of its other Tfp chemotaxis components [50]. The knock out results are consistent with the absence of CheB and CheR proteins from the majority of Tfp systems, and this information challenges previous predictions that bacteria are incapable of spatial sensing due to their small size [191] as will be discussed further in Chapter 3.5.2. Both Alt and flagellar families utilize typical CheB proteins made up of a REC and methylesterase domain, but the Alt system is associated with CheR-TPR fusion proteins described in Chapter 3.2.2.1. In *E. coli*, CheB and CheR proteins are localized to the sensory lattice via a small pentapeptide that is attached to some MCPs by a flexible loop [152,192]. The pentapeptide, and thus this method of localization, is only present in a subset of organisms [34]. This makes it likely that other chemotaxis systems may have developed different methods of CheB and/or CheR localization much like P2-Hpt domain of Tfp systems in contrast to the standard flagellar P2 domains.

We have been able to identify the same three functional families and 10 flagellar classes in the CheB/CheR (Figure 3.9) and CheA trees (Figure 3.7) and confirm them by gene neighborhood analysis since at least one member of each CheB/CheR pair is encoded near CheA in 172/189 of the chemotaxis associated pairs. However, members of some CheA subfamilies lack CheB and/or CheR, and, even rarer, some CheBR subfamily members lack an associated CheA. These discrepancies are largely due to the partial lateral transfer of a system and/or system degradation based on phylogenetic and gene neighborhood analysis. The Tfp systems of the three *Xanthomonas* species and

Xylella fastidiosa contain CheB, but they lack CheR and are assumed to be non-functional due to extreme divergence. In contrast, the Flb systems of *Listeria innocua*, *L. monocytogenes*, *Bacillus thuringiensis*, *Bacillus cereus*, and *Bacillus anthracis* have only CheR, but their CheR sequences do not show significant divergence from closely related F1a systems that have retained CheB. It is unclear whether or not they participate in the

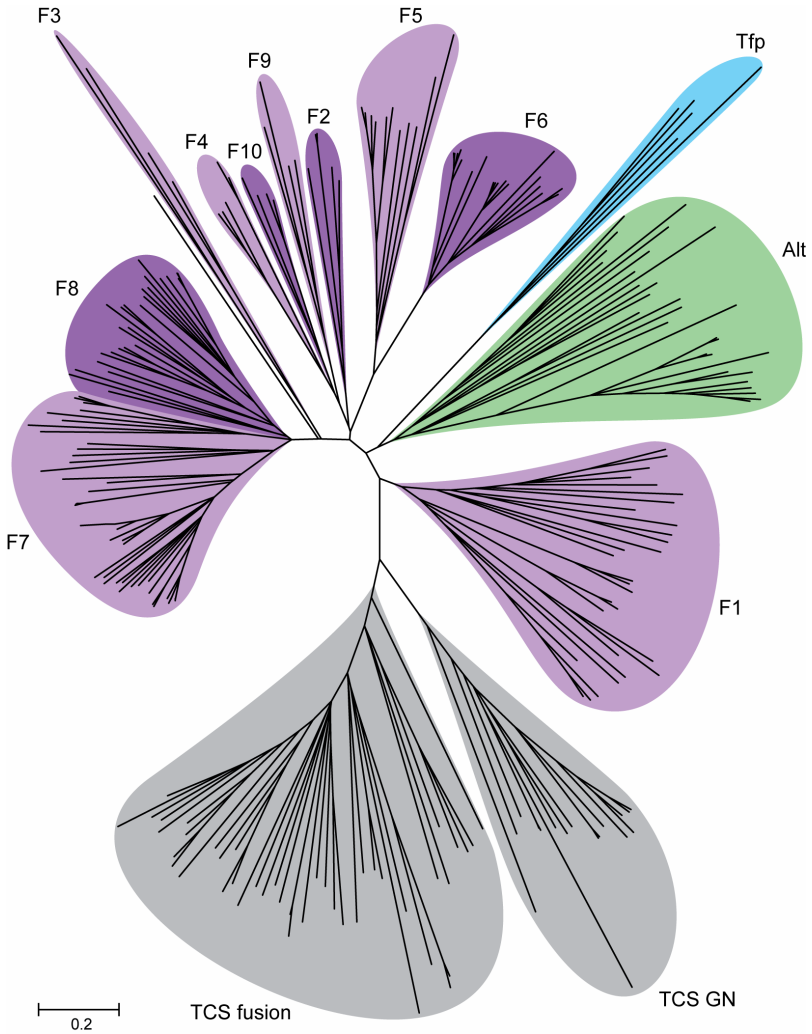


Figure 3.9 Phylogenetic analysis of a concatenated alignment of CheB and CheR protein pairs shows the same subfamilies identified in Figure 3.7 in addition to two subfamilies associated with HKIs. The TCS fusion family contains CheB and CheR pairs that are typically fused together and often fused to HKI signalling modules. The TCS GN group is associated with TCSs based on gene neighborhood data. The members of each group can be found in Tables A.3 and A.4.

chemotaxis system, but studies in *L. monocytogenes* support that it is capable of chemotaxis [148,166]. *H. pylori* has been experimentally shown to be chemotactic [145], but its chemotaxis system and that of *Helicobacter acinonychis* lack CheB and CheR despite their presence in the remaining members of the F3 subfamily. The F3 CheB proteins lack the N-terminal REC domain, but the F3 subfamily is also the only flagellar system that utilizes a CheAY fusion protein. It is possible that the REC domain of F3 CheAs came from the CheBs of these systems and has gained a new function given that the CheA REC domain of *H. pylori*, which lacks CheB, has been shown to be phosphorylated in vitro [193]. A few members of the F6 group lack associated CheB and CheR proteins, but their systems are presumed to be incomplete due to reliance on, and possible interaction with, F7 systems that are also encoded in the organisms. Evolutionary factors that are predicted to have shaped the distribution of CheB and CheR will be discussed in Chapter 3.5.2.

3.3.2 CheD Analysis

CheD plays a role in the adaptation pathway by deamidating key glutamine residues of MCPs into glutamate residues so they can be methylated by CheR [35]. In addition to its role in deamidation, CheD is also involved in increasing CheC's phosphatase activity on phosphorylated CheY (CheY-P) [29,36]. The phyletic distribution of CheD and CheC shows that many organisms that have CheD lack CheC [194]. In *E. coli*, CheB is responsible for de-methylation and deamidation [195]. Many other organisms have CheB and CheR but lack CheD, like *E. coli*. The domain architecture reveals that all CheD proteins are single domain proteins with the exception of a CheD-CheB fusion found in *B. bacteriovorus*. A multiple alignment of CheD sequences shows that all members contain conserved residues predicted to be involved in catalytic activity except for one extremely divergent sequence found in a virus, *Acanthamoeba polyphaga mimivirus*, that not surprisingly lacks all other chemotaxis

components and was removed from the final analysis. Gene neighborhood analysis shows that the majority of CheD proteins are encoded in the genomes near other chemotaxis proteins, implicating their involvement in chemotaxis regardless of the presence of CheC and making it easy to correlate them with chemotaxis subfamilies (Figure 3.10).

Gene neighborhoods show that CheD is almost exclusively involved in F1, F7, and F8 chemotaxis systems. Only the F1 CheD sequences are associated with CheC interaction based upon gene neighborhood and mirror tree analysis. The F7 and F8 systems have been shown to be closely related in the CheA and CheBR trees (Figures 3.7 and 3.9), and the shorter sequence length and smaller data set for CheD analysis decreases the resolution of the F7 and F8 classes. The CheD tree shows that the CheD proteins of the older F7 systems from ε and δ-Proteobacteria group with the F8 systems instead of the F7 systems of α and β/γ-Proteobacteria. Gene neighborhood data and CheA phylogeny already show a close relationship between the F7 and F8 systems (Figure 3.7), The F8 system from *Dechloromonas aromatica* has two CheD proteins encoded in its gene neighborhood; one is associated with the F8 system while the other groups with the F7 system. Surprisingly, three of the older F7 CheD sequences group more closely with the F1 class, which is most likely due to the somewhat poor tree resolution. The F7 CheD proteins from α, β, and γ-Proteobacteria contain a C-terminal extension, and the VISSA program [110] reveals that the extension forms a distinct structure with a conserved α-helix followed by a possible β-strand. The β-strand caps the end of most of the CheD members of this subfamily and bears sequence similarity to the NWETF pentapeptide that is associated with binding CheB and CheR to MCPs that have the sequence [34,152,192]. The CheD pentapeptide has a loose consensus of X-[ILV]-[polar]-[ILV]-[F]. The similarity suggests that the CheD pentapeptide may also play a role in CheB and CheR interaction to localize the two proteins to the signaling complex. The CheD pentapeptide is only found in systems that have MCPs with pentapeptides. It

is possible that CheB and CheR may have different pentapeptide binding preferences to reduce competition, or such an interaction may allosterically affect the catalytic activity of any of these three enzymes.

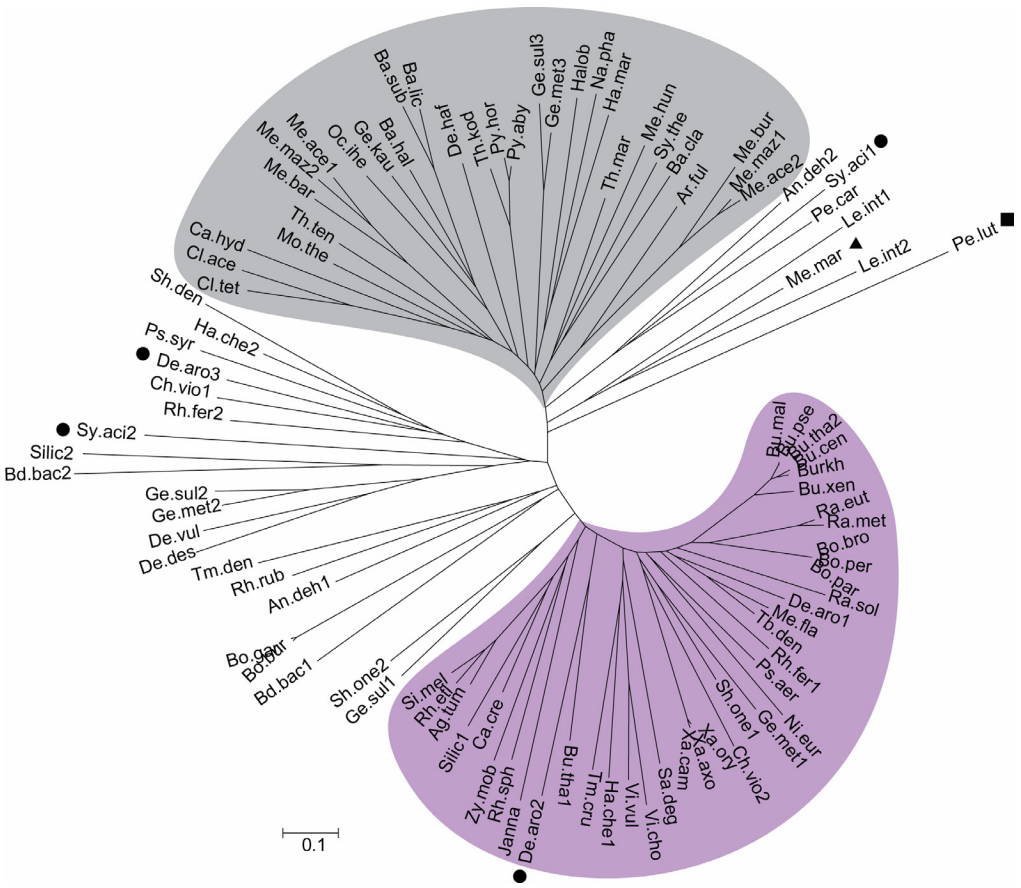


Figure 3.10 Phylogenetic analysis of CheD sequences shows poor subfamily grouping. The F1 associated CheD sequences group together (shown in gray) with the exception of the CheD from *Methanococcus maripaludis* (marked by a triangle). The F1 sequences are predicted to interact with CheC based on gene neighborhood data. F7 and F8 CheD sequences do not form distinct groups. A subset of F7 CheD sequences group together (shown in purple) and have a C-terminal extension that resembles the pentapeptide motif associated with MCP-CheR [19] and MCP-CheB [196] interactions in *E. coli*. Circles mark two pairs of sequences that are encoded in the same gene neighborhoods of *D aromatica* (De.aro) and *Syntrophus aciditrophicus* (Sy.aci), but show different groupings. The De.aro pair is encoded near an F8 CheA, and the Sy.aci pair is encoded near an F7 CheA. The square marks a CheD found in the organism *Pelodictyon luteolum* that lacks all other chemotaxis components. Sequence identifiers correspond to Table A.5.

3.3.3 CheZ Analysis

CheZ is a single domain protein that dephosphorylates CheY-P [25,153]. The CheZ similarity search revealed homologues in δ and α-Proteobacteria although the CheZ domain (Pfam:CheZ) was absent from these members and CheZ has never been reported in either of these taxonomic classes. Recently CheZ was experimentally identified in a member of ε-Proteobacteria [197] after previous claims of its absence [198,199], which was confirmed by our finding members of ε-Proteobacteria in CheZ similarity searches. A multiple alignment shows that these divergent CheZ proteins do contain the conserved catalytic glutamine residue (Gln 147 in *E. coli*) for CheY-P dephosphorylation [25] and some are encoded near CheY-like proteins. The phylogenetic tree shows subfamilies that can be clearly correlated to the F3, F4a, F5, F6, and F7b chemotaxis subfamilies of the CheA tree based on gene neighborhood and mirror tree analysis. The experimentally characterized CheZ of *E. coli* is found in the F7b subfamily. The *E. coli* CheZ interacts with CheY-P and CheA at seperate regions that are both found in all members of the F6 and F7b subfamilies except for the F6 systems of *Thiomicrospira crunogena* and the three *Xanthomonas* species, which lack the CheA binding region [200]. CheZ interacts with CheY at two distinct regions identified in the CheY-CheZ co-crystal structure, which are conserved in the CheZ alignment [25,30]. The catalytic region interacts with the primary face of CheY for direct CheY-P dephosphorylation. A small C-terminal alpha helix of CheZ (CheZc) interacts with the α4-β5-α5 face of CheY that has also been implicated in motor interactions with FliM [201-203].

Enterobacteria CheZs group more closely with β-Proteobacteria than with γ-Proteobacteria, which supports the previous hypothesis that Enterobacteria obtained their chemotaxis system by lateral transfer as shown in the CheA analysis. The F3 associated CheZ proteins have an elongated CheA binding region that shares no sequence similarity with the F6/F7b CheA binding region, but the existence of this subdomain suggests that it

may still be involved in binding CheA. Although the F3 CheZ sequences are not found encoded with other chemotaxis components, the CheZ of *H. pylori* has been shown to play a role in chemotaxis [197]. The F4a and F5 subfamily CheZ sequences lack the CheA-binding region and are not predicted to interact with CheA. The CheZ of F5 systems are typically not encoded near chemotaxis components, but the CheZ of *C. crescentus* is divergently transcribed from the F5 chemotaxis system (see Ca.cre2 in Table A.2), which further supports the involvement of this subfamily in chemotaxis. The phyletic distribution shows that CheZ originated in a chemotaxis system of the common ancestor of Proteobacteria.

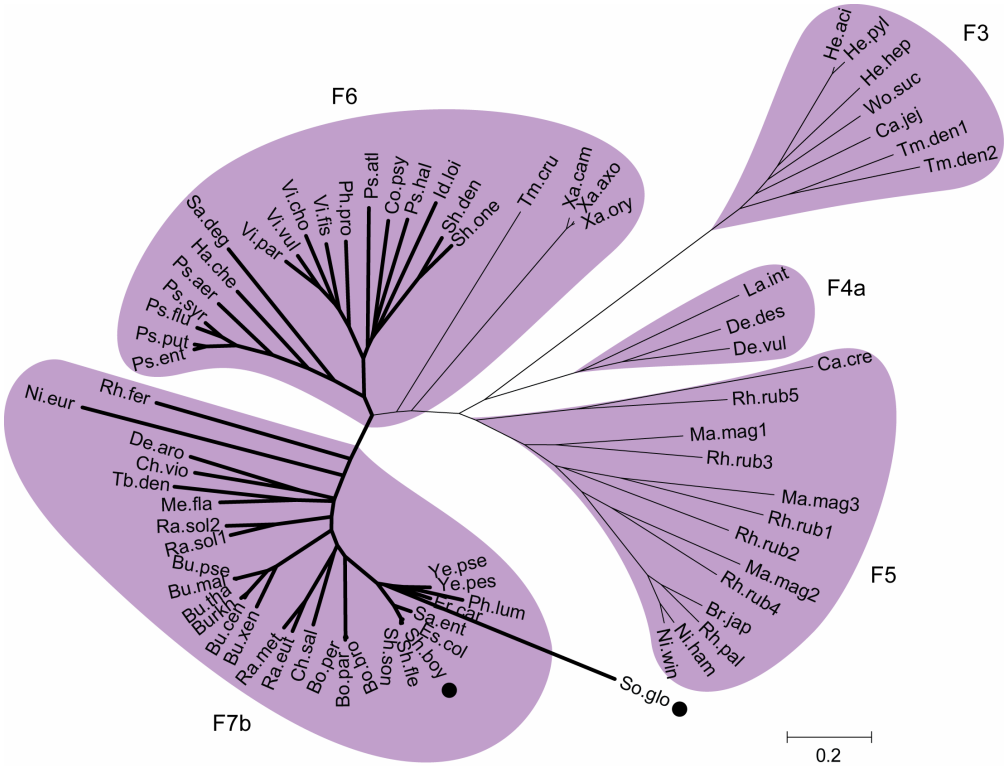


Figure 3.11 CheZ sequences group together based on chemotaxis subfamilies and structural differences. CheZ sequences are found in F6 and F7a gene neighborhoods. Phylogenetic profiling [78] is used to link CheZ sequences to the F3, F4a, and F5 subfamilies. Black circles mark CheZ sequences from organisms that have lost their associated CheA. Bold lines indicate CheZ proteins that have the experimentally identified CheA-binding region [200]. Sequence identifiers correspond to Table A.6.

3.3.4 CheC and CheX Analysis

The crystal structures of the closely related CheC and CheX proteins have been solved revealing distinct differences in their structures and interactions [27,29]. The CheCYX family of CheY phosphatases share similar sequences but different structures and domain architectures [121]. CheC and FliY have two homologous active sites (Pfam:CheC), but FliY has an additional C-terminal domain (Pfam:SpoA) that is involved in structural assembly with the exception of the split FliY in *T. maritima* [121]. FliY is located at the flagellar motor whereas CheC is predicted to be localized to the MCP lattice via CheD, which activates CheC mediated dephosphorylation of CheY-P. The CheX phosphatase is closely related to CheC, but it has only the second active site of CheC (Figure 2.1) and exhibits differences in structure at the secondary and tertiary levels. The major site of structural differences between CheC and CheX corresponds to their CheD interaction and dimerization sites, respectively [27,29]. CheC and CheX are poorly conserved, and many homologues identified by similarity searches lacked identifiable CheC domains. A multiple alignment revealed that the sequences do show CheC/CheX active site residues [27], revealing the Pfam:CheC domain model to be incomplete. Although CheC has been characterized as having two active sites, the multiple alignment also shows that some CheC proteins lack the catalytic residues at one active site. This is most commonly found in organisms that have multiple CheC proteins encoded in their genomes (Figure 3.12). The FliY proteins were removed from the final alignment as they are part of the flagellum more so than the chemotaxis system. The VISSA program for secondary structure analysis of a multiple alignment was able to help distinguish the CheC and CheX subfamilies which were then realigned separately to yield two trees.

Gene neighborhood and domain architecture data show that CheC is almost always encoded near or fused to a CheY protein. CheC phylogenetic analysis shows two distinct subfamilies. One is associated with F1 systems, all but two of which are

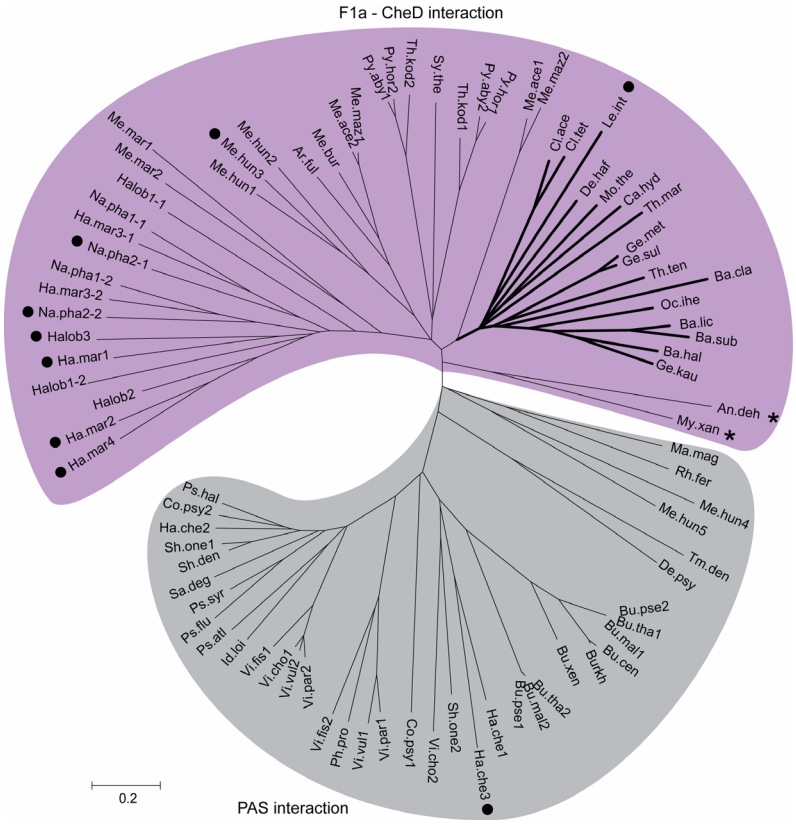


Figure 3.12 All CheC proteins are predicted to interact with CheY, but two CheC subfamilies show additional associations. Gene neighborhood data links one family with F1a systems and CheD, a known partner of CheC. The other family is linked to TCS proteins that contain PAS domains [185]. Black circles represent sequences that are not encoded near CheD or PAS within each subfamily. Asterisks identify the CheC of F1 systems that were laterally transferred into some δ-Proteobacteria. Sequences with CheC-CheC fusions are identified by the sequence identifier and a 1 or 2 corresponding to the first or second CheX region in the protein, respectively. Despite the low sequence conservation of CheC, the tree still predominantly shows taxonomic based groupings as seen with the bold branch group that contains most of the Firmicute CheC sequences. Sequence identifiers correspond to Table A.7.

predicted to interact with CheD based on gene neighborhood and mirror tree data. The *M. xanthus* F1 system was received by lateral transfer and has CheC (My.xan in Figure 3.12), but lacks CheD. Experimental studies have shown that the CheC protein still plays a role in *M. xanthus* [204]. The other CheC subfamily members are not linked to specific chemotaxis systems, but all are encoded immediately adjacent to a protein that contains a PAS domain and a variable output domain. This tight correlation suggests that CheC or

its associated CheY may be able to interact with the PAS sensory domain in order to regulate the activity of CheC and/or the PAS associated output domain since these CheC proteins are not associated with CheD interaction. However, CheC has only been identified in organisms that contain chemotaxis systems, which also supports its involvement in chemotaxis even if indirect. CheC domain architecture analysis reveals CheC-CheC and CheA-CheC fusion proteins in addition to CheY-CheC fusion proteins. The CheX protein from the F2 chemotaxis locus in *B. burgdorferi* has been shown to be essential for chemotaxis [155] and is present in all members of the F2 class. The CheX protein is not consistently associated with any other chemotaxis classes. Three members of the family are encoded in or near F7 system gene neighborhoods of *Pelobacter carbinolicus* and *Syntrophus aciditrophicus*, and two are encoded near the F5 system neighborhood of *Desulfotalea psychrophila* (Figure 3.13). The F2 associated CheX proteins group tightly with CheX proteins from organisms that have F1 systems, although CheX is lacking in most F1 systems. Given its general role and limited interactions, CheX may be able to be more successfully laterally transferred or more easily lost by chemotaxis systems. CheX has been shown to dephosphorylate CheY-P *in vitro* in *T. maritima* [27], which has an F1 system, but whether or not CheX plays a role in chemotaxis in all organisms that have it or is a phosphatase of REC domains in other signal transduction systems is currently unknown. CheX is exclusively found in organisms with flagellar chemotaxis systems, which does support a continued role in chemotaxis. The CheX multiple alignment reveals a small subset of CheX proteins with a C-terminal extension that is predicted to end in an α-helix that is very similar in sequence to the CheZc α-helix known to bind CheY [25,30]. Domain architecture analysis shows CheX predominantly exists as a single domain protein, but there are some CheY-CheX and CheX-CheY fusion proteins.

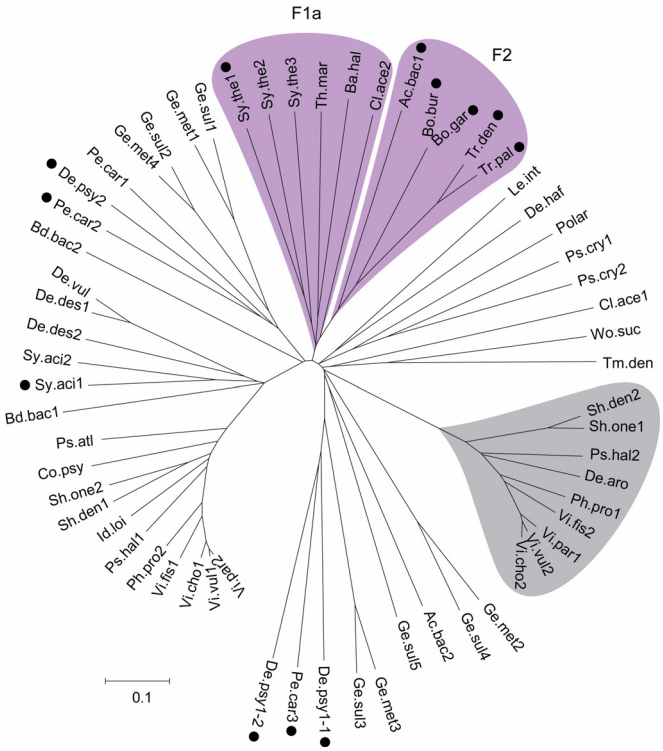


Figure 3.13 CheX proteins are not highly correlated to flagellar chemotaxis classes except for F2 systems. Black circles mark CheX proteins that are encoded near *cheA*. The F2 CheX proteins are encoded near the F2 *cheA* and have been experimentally shown to aid chemotaxis [155]. F2 CheX proteins group with F1 associated CheX proteins, but the majority of F1 systems lack CheX. A subfamily of CheX proteins have a C-terminal extension that shows sequence similarity to the C-terminal region of CheZ that is associated with CheY interaction [25,30]. Sequences with CheX-CheX fusions are identified by the sequence identifier and a 1 or 2 corresponding to the first or second CheX region in the protein, respectively. Sequence identifiers correspond to Table A.8.

3.3.5 CheV and CheW Analysis

CheW is involved in sensory lattice scaffolding by interacting with CheAs and MCPs [205]. The CheW protein is a single domain (Pfam, SMART:CheW) that is homoglous to domains in all CheV and CheA proteins in addition to an unusual CheW-CheR fusion protein exclusively found in F2 systems. The CheW protein is associated with all chemotaxis systems with the exceptions of *L. innocua*, *L. monocytogenes*, *B. cereus*, *B. anthracis*, and *B. thuringiensis*, which exclusively use CheV in place of CheW. CheW proteins are typically found in major chemotaxis loci with CheA. In small but

significant portion of these loci, there are duplicate CheW proteins. On a phylogenetic tree the duplicate CheWs often fall into different subgroups based on significant sequence divergence, which raises questions as to the function of these multiple CheW proteins. Similarly there are also a few CheW proteins that contain multiple CheW domains, and these domains have diverged significantly based on sequence. Although the multiple CheW domain proteins exist outside of the F9 subfamily, all F9 systems have a protein with three CheW domains encoded in their gene neighborhoods rather than a single domain CheW protein. The MCP found in F9 neighborhoods is cytoplasmic and consists solely of two copies of the signaling module. The other proteins with three CheW domains are typically found encoded next to a cytoplasmic MCP, which suggests that multidomain CheW proteins may aid in stabilizing a higher order chemotaxis complex in the cytoplasm of some organisms. In addition to CheW proteins with multiple CheW domains, domain architecture also revealed CheW proteins fused to PAS [185] or Cache [206] sensory domains. A PAS-PAS-CheW fusion protein and a few CheW proteins with long undefined N-terminal extensions that were identified as Cache-Cache-CheW proteins after PSI-BLAST searches of the undefined regions. The CheW domain is poorly conserved in comparison to most of the chemotaxis proteins due to its lack of enzymatic function. Although subfamilies on the CheW tree can be identified, the overall tree topology is very poor and the more indepth sequence analysis that is needed to understand the functional differences in these proteins is outside the scope of this study.

CheV contains a CheW domain that is implicated in similar interactions as CheW [154] and a C-terminal REC domain that is regulated via phosphorylation by CheA [37,198]. Phylogenetic profiling shows that CheV proteins are associated with F1, F3, F4, F6, and F7b subfamilies (Figure 3.14). Although our analysis shows F1 CheVs are almost exclusively present in Bacilli, BLAST searches against incomplete Firmicutes genomes show it to be present in multiple members of Clostridia, which supports that

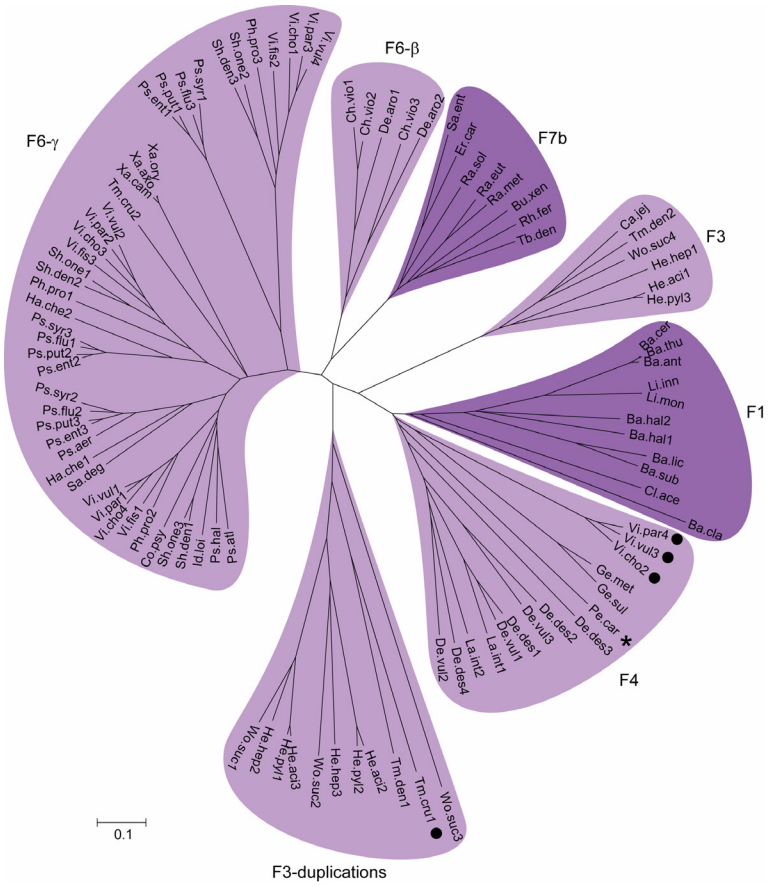


Figure 3.14 There are multiple CheV duplications in the F3 and F6 chemotaxis systems. Although only one F3 group contains all F3 CheA members, only He.pyl1 in the F3-duplications group has been shown to be necessary for chemotaxis in experimental studies [198]. Black circles mark laterally transferred CheV sequences that lack similar associated components, and an asterisk marks a CheV sequences that is hypothesized to have lost its cognate F4 components. Sequence identifiers correspond to Table A.9.

CheV has been vertically propagated to these diverse subfamilies. CheV proteins usually are not encoded near other chemotaxis proteins, although they sometimes are found near flagellar proteins and near CheR and always are encoded next to F3 CheAs. Specific subfamilies and duplication events can be identified, but there are no sequence features that can be related to functional differences between the subfamily. Since both CheW and CheV have primary roles in scaffolding with fewer dynamic interactions than other chemotaxis proteins, it is possible that evolutionary pressures on these components have

resulted in more divergence at the individual sequence level rather than insertions and deletions that can be more easily traced and identified.

3.3.6 CheY

The single domain CheY protein consists of the REC domain that is an essential signalling module of TCSs [115]. Although the majority of TCS associated proteins with REC domains can be distinguished by the presence of additional output or signalling domains, stand alone REC domains can also serve as middlemen in extended TCS phosphotransfer relays [13,128]. We used sequence similarity searches with length filters and subsequent phylogenetic analysis to attempt to distinguish true CheY proteins from other stand alone REC domains (Chapter 2.3.2.4). Our data set probably includes some stand alone REC domains, but the majority can be confidently predicted as CheY proteins based on gene neighborhood data. Our initial REC domain analysis identified two families encoded in chemotaxis gene neighborhoods. Members of one subfamily are encoded in Tfp chemotaxis gene neighborhoods, and members of the other subfamily are associated with flagellar chemotaxis systems. Most Tfp systems have two CheY proteins in their gene neighborhoods, but one of the CheY proteins in the cyanobacterial Tfp systems has an elongated N-terminal domain [207]. These unusual proteins are homologous to a protein called PatA that has been shown to affect heterocyst formation in *Nostoc* sp. 7120 [208]. A mutation in one of the PatA-like CheYs from *Synechocystis* sp. results in chemotaxis defects [63]. Both families are included in our CheY data set. Phylogenetic analysis of the final set confirms the Tfp family and also shows groupings associated with the flagellar subfamilies. Although the CheY protein is more highly conserved than the CheW protein, many systems have paralogous CheY proteins that obscure the phylogenetic analysis.

The CheY tree topology reflects its phosphotase interactions, an observation that will be exploited for detailed molecular evolutionary analysis in Chapter 5. One

subfamily is found to be associated with CheZ interaction based on gene neighborhood and CheZ phyletic distribution. Members of another subfamily are associated with CheC and CheX interaction based on gene neighborhoods. Given that some CheC members have N-terminal REC domains instead of neighboring CheY proteins (Chapter 3.3.4), the REC domains fused to CheC were manually added to the CheY alignment. The subsequent phylogenetic analysis shows that these REC domains clearly group with the CheC associated CheY proteins. Although these proteins have domain architecture more typical of a TCS component, it is possible that they have CheY-like functionality with a cis-acting phosphatase domain.

3.4 Flagellar Subfamily Characterization

The expanded component repertoire of flagellar chemotaxis systems shows that they are far more diverse than Tfp and Alt systems in addition to being more numerous as supported by the CheA phylogenetic tree (Figures 3.2 and 3.7). Flagellar subfamilies identified in the CheA and CheBR analyses can be clearly correlated to eight of the 12 MCP length classes [34] by phylogenetic profiling [132]. The distribution of the chemotaxis families and classes in comparison to MCP length class distribution and their correlations are shown in Figure 3.15. All of the organisms that have an MCP class but no associated CheA have only one member of the MCP class. The three lone F6 type 40H MCPs are predicted to be the result of lateral transfer and lack significant chemotaxis system involvement. The 36H MCP of *Wolinella succinogenes* groups with the 36H MCP of closely related *Thiomicrospira denitrificans* in a 36H tree and is predicted to have lost its associated F7 CheA. Organisms with a CheA that lack its cognate MCP are predicted to be systems in the process of being lost with the exception of *Symbiobacterium thermophilum*. This organism is associated with the 52H MCP class, which is exclusive thus far to it and is predicted to have resulted from an insertion in the 44H class typically associated with F1 systems. Like *S. thermophilum*, the F5 system of

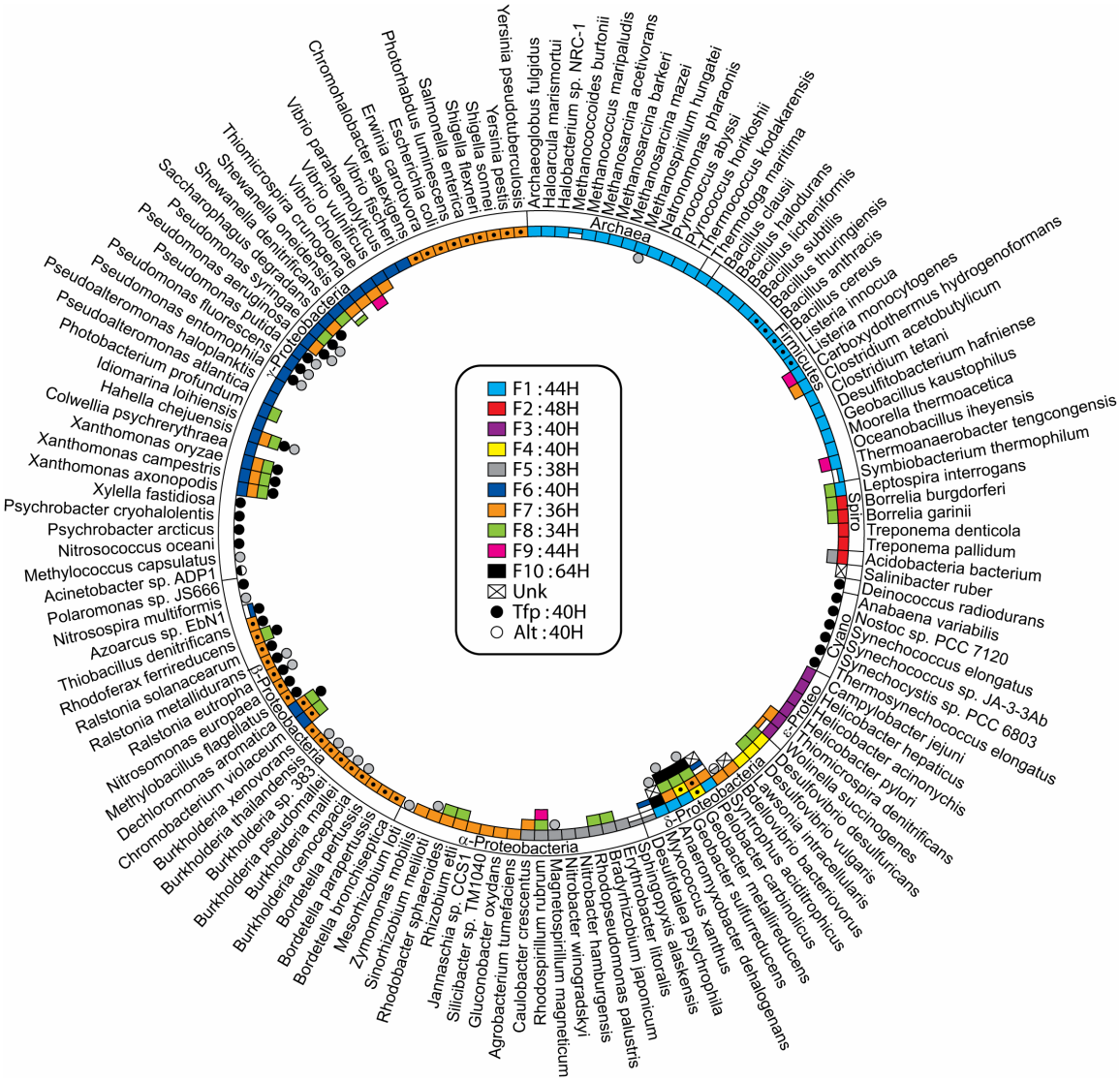


Figure 3.15 Chemotaxis system distribution and correlation of the classes with eight MCP length classes. Although the 40H MCP class is associated with multiple chemotaxis families, the families can be clearly distinguished on a phylogenetic tree of the 40H MCPs (Table A.12). In addition to 40H MCPs, all F3 systems are correlated with 28H MCPs (Table A.13). Automated MCP classification did not reveal any length class associated with the F9 family, but manual analysis of MCPs in the F9 gene neighborhoods revealed them to be 44H members with nominal gaps that excluded them from identification by the MCP class HMMs. Organisms that have only the CheA associated with a family have the outermost portion of the block or circle colored, and organisms that have only the MCP class and not the associated CheA are colored in the innermost portion of the circle. The MCPs identified in this study can be found along with their automated classifications in Table A.13.

D. psychrophila lacks the associated 38H MCPs due to a 4 heptad insertion that resulted in its unique 42H MCP class. The 24H MCP class is not shown since it shows no correlation to specific chemotaxis classes. No 58H members were found in this genome set. The P2 domain as defined by previous experimental work is found only in flagellar systems and has three related subtypes (Figure 3.6), two of which have been crystallized revealing structural differences. We determined that CheB and CheR are associated with the vast majority of flagellar chemotaxis systems, and are thus considered core components along with CheA, CheW, CheY, and MCPs unless otherwise stated. Phylogenetic analyses of each accessory component (CheC, CheD, CheV, CheX, and CheZ) link them to specific subfamilies by mirror tree analysis in addition to standard gene neighborhood analyses. All of this information allows us to clearly characterize each subfamily and to reveal insights into each of their mechanisms.

3.4.1 The F1 system

The chemotaxis proteins from Archaea and Firmicutes typically group together in phylogenetic analyses of each component, forming the F1 (flagellar type 1) class which includes two subtypes. All members of Archaea have F1a systems as do most Firmicutes. Archaeal F1 systems form two groups in the F1 family of the CheA tree, but have a distinct clade within the F1 group of the CheBR tree. Their distant phylogenetic relationship supports the hypothesis that the common ancestor of Euryarchaeota received a chemotaxis system via lateral transfer from a member of Firmicutes. Chemotaxis in Gram-positives is exclusive to Firmicutes (low GC Gram-positives) including *S. thermophilum*, an unusual bacterium that has a high GC content like Actinobacteria but is a member of Firmicutes based on whole genome analysis [209] and our 16s analysis. The F1a system has been studied extensively in *B. subtilis* [33] and *H. salinarum* [146,210], but the unusual F1b system has also been shown to be important for motility in *L. monocytogenes* [148,166] despite its origin via lateral transfer (Figure 3.2). All

members of this family utilize 44H MCPs and have a P2-I domain within CheA, except the F1a system of *S. thermophilum* that has had an insertion in its 44H MCPs resulting in 52H MCPs [34]. F1a systems utilize the accessory proteins CheC and CheD. The CheX and CheV accessory proteins are present in some F1a systems, but are not essential to all. F1b looks like a reduced version of F1a in that it always has CheV, but lacks CheC, CheD, CheX, and surprisingly CheW and CheB. The purpose of the remaining CheR in this system is not clear, but all of the members do retain their catalytic residues and lack excessive divergence. An F1a system is present in the spirochete *L. interrogans*; it is significantly divergent based on the CheA tree, but it is associated with P2-I domains, 44H MCPs, CheC, and CheD like F1a systems (although the CheC is fused to CheA). It is not clear whether or not this system was received by lateral transfer or horizontal inheritance. An F1a system was laterally transferred to a common ancestor of some δ-proteobacteria, but significantly diverged within the organisms, thus eliminating the expected groupings within the CheA tree despite the fact that all have retained the 44H MCPs and some have retained CheC and/or CheD. The Dif system of *M. xanthus* is one of these unusual F1a systems [59,171,204].

3.4.2 The F2 System

The F2 chemotaxis systems are found exclusively in Spirochetes and have been studied in *B. burgdorferi* [155] and *T. denticola* [149,167]. They contain all of the core proteins and are characterized by 48H MCPs, a CheA with P2-I, a CheW-CheR fusion protein, and CheX. A divergent F2 system that lacks the CheW-CheR protein is also present in *Acidobacteria bacterium*, presumably by lateral transfer (Ac.bac2 in Figure 3.2) since this organism groups with δ-Proteobacteria in a universal protein family tree [211]and in our 16S rRNA tree (Figure 3.3). The MCP in its gene neighborhood does not match the 48H HMM (nor any other length class), but CheBR (Ac.bac2 in Table A.3) and CheX (Ac.bac1 in Figure 3.13) trees support its grouping with the F2 family.

3.4.3 The F3 System

F3 chemotaxis systems have been found in all sequenced ε-Proteobacteria. They all contain the core proteins with the exception of the closely related *H. pylori* and *H. acinonychis* which have lost both CheB and CheR. The CheBs of those that retain the methylation system lack the N-terminal REC domain. *H. pylori* has nonetheless been shown to have a functional chemotaxis system [145]. A possible explanation for its lack of CheB and CheR can be found through analysis of the evolution of the system (Chapter 3.5.2). F3 systems use 40H and 28H MCPs (the one 40H MCP in *Campylobacter jejuni* was missed by HMM analysis) and utilize the accessory components CheV (often multiple copies as shown in Figure 3.14) and CheZ. The CheZ has an extension corresponding to the CheA-binding domain experimentally identified in *E. coli*, but the multiple alignment shows that it shares no sequence homology. The CheZ protein of *H. pylori* has been shown to be essential for normal chemotaxis [197], but in depth analysis has not yet shown if it binds CheA. The F3 CheA lacks the P2 domain, but does have a C-terminal REC domain. It seems plausible that the REC domain could be the missing REC domain of CheB despite the loss of the remaining catalytic portion of CheB in some species, but phylogenetic analysis has not been able to clearly support or eliminate this hypothesis. Regardless, in vitro studies have shown the CheA REC domain to be phosphorylated [193].

3.4.4 The F4 System

The F4 system is found in δ-proteobacteria with two different subtypes. Type F4a is in *Lawsonia intracellularis* and both *Desulfovibrio* species, while Type F4b is in both *Geobacter* species. Both F4 types lack P2 domains and use 40H MCPs and CheV. F4a CheAs have three HPT domains, and their systems are associated with CheZ proteins by mirror tree analysis and with HEAT repeat proteins (SMART: EZ_HEAT) by gene neighborhood data. An alignment of CheZ proteins shows that those associated with F4a

systems lack the CheA-binding domain. The F4b systems contain two CheR proteins and a hydrolase (SMART: HDc) encoded within their gene neighborhoods and lack an encoded CheY protein. The second *cheR* gene in the locus encodes a more divergent protein. A candidate receiver is the sigma (54) type response regulator divergently transcribed upstream from the F4b locus (though split by a transposon in *Geobacter sulfurreducens*) since it has an N-terminal REC domain, plus a similar divergently transcribed gene encodes a protein that has been found to be regulated by a chemotaxis system in *M. xanthus* [54]. The unclassified laterally transfered chemotaxis system of the Bacteriodetes member, *S. ruber*, groups with F3 and F4b systems in the CheA tree and like both families it has 40H MCPs and lacks P2 domains. Unlike P3 and P4 systems, this system lacks a CheV protein, and the CheB and CheR proteins (found encoded near its CheA) showed very different groupings in the individual trees, which resulted in their exclusion from the concatenated CheBR alignment and phylogenetic analysis. The similarity of the *S. ruber* chemotaxis system to both F3 and F4 systems suggest that it may have come from a lateral transfer event of their common ancestor that has since diverged.

3.4.5 The F5 System

F5 systems are found primarily in α-proteobacteria, but they are also present in *A. bacterium* and *D. psychrophila*. The majority of the F5 CheA proteins are CheAV fusions. The system has been shown to be essential for chemotaxis in *A. brasilense* [61] and *R. centenum* [60]. Although four of the CheAV proteins lack the C-terminal REC domain, they are from divergent systems that are predicted to be in the process of being lost from their associated genomes. Despite a few domain architecture inconsistencies, CheAV fusion hypothesis is supported by system evolution analysis (Chapter 3.5.3). These systems are associated with 38H MCPs and CheZ has been associated with all of the α-Proteobacterial systems except for the very divergent systems in *E. litoralis* and

Sphingopyxis alaskensis (see the long branches of Er.lit and Sp.ala in Figure 3.2). Like the F4a system, the F5 CheZs also lack the CheA-binding domain (Figure 3.11).

3.4.6 The F6 and F7 Systems

The F6 systems are primarily found in γ-Proteobacteria, although there are a few incomplete systems in β-Proteobacteria. All F6 systems have CheV and CheZ, but the β-Proteobacteria F6 systems lack CheB and CheR, as do the F6 systems from the three *Xanthomonas* species and *T. crunogena* of γ-Proteobacteria. F6 systems utilize 40H MCPs, and the F6 CheAs in γ-Proteobacteria typically have a conserved repeat within the region that corresponds, but has no sequence similarity, to the P2 domain region, which leads us to predict that this repeat could be involved in CheY interaction. F6 systems have a sordid evolutionary relationship with F7 systems. F7 systems are found in all classes of Proteobacteria and can be categorized into two types. All F7 are associated with 36H MCPs and CheD. The majority of F7a CheAs have a P2-II domain, but it is absent in δ and some of the γ-Proteobacteria members. The F7b systems are the result of a fusion of the F7 and incomplete F6 systems in β-proteobacteria. They use all of the F7a components and have simply adopted the CheV and CheZ of F6 systems, although CheV has been lost in most of these systems (Figure 3.14). The addition of CheZ resulted in the reduction of the P2-II domain into a P2-III domain, like the one of *E. coli*. Although *E. coli* is a member of γ-Proteobacteria, its chemotaxis system components group tightly with β-Proteobacteria F7b systems. It seems likely that the *E. coli* system was received laterally from a member of *Bordetella* since it is the only β-proteobacterial F7b member that lacks a CheD encoded in its major chemotaxis locus (although it is present elsewhere in the genome: Bo.bro, Bo.per, and Bo.par in Figure 3.10) and all of the F7b members found in γ-Proteobacteria lack CheD entirely. It is also worth noting that the large majority of F7 CheDs in α, β, and γ-Proteobacteria have a C-terminal extension that

contains a motif very similar to the MCP pentapeptide motif that has been shown to be important for binding CheB and CheR in *E. coli*.

Whereas most F7b systems have resulted in the loss of the F6 CheA, *Chromobacterium violaceum*, *D. aromatica*, *T. crunogena*, and the three *Xanthomonas* species contain CheAs from both F6 and F7 although their F6 systems are incomplete. In the first organisms, β-Proteobacteria, the two loci are encoded adjacent to each other within the genome, and which supports that the two systems work together. There are many 40H MCPs associated with the F6 system and only one 36H MCP associated with the F7b system; this implies that the 40H MCPs were the main sensors in the two organisms before the 36H MCPs later became predominant. The γ-Proteobacteria *T. crunogena* and *Xanthomonas* species lack any 40H MCPs (Figure 3.15), and instead have undergone massive lineage specific expansion of the 36H MCPs (Table A.13 and phylogenetic analysis not shown). The F6 and F7 systems of these six organisms seem to represent a snapshot of what the flagellar chemotaxis systems looked like in the common ancestor of β/γ-Proteobacteria before they fused in β-Proteobacteria and differentiated in γ-Proteobacteria. Some γ-Proteobacteria have complete F6 and F7 systems, but in these cases F6 systems seem to be dominant based upon a higher number of MCPs associated with the systems. Additionally all γ-Proteobacteria with flagellar chemotaxis systems have F6 systems except for those that received their systems by lateral transfer, like *E. coli*. Experimental studies in *P. aeruginosa* that show the overexpression of the F7a CheB restores the system to near wild-type levels when CheB of the F6 system (the major flagellar chemotaxis system) is deleted [62], which supports the potential for crosstalk between the systems *in vivo*. However, more recent studies in *P. aeruginosa* show that the F7a system components do not co-localize with the F6 system components in vivo, and the one 36H MCP encoded in the genome is required for the localization of the F7a components [41]. This study supports our prediction that these systems primarily act independently when both contain complete component repertoires.

3.4.7 The F8 System

F8 systems are present in δ, α, β/γ-Proteobacteria, and Spirochetes, and their components consistently group with F7 components, which supports the hypothesis that the two systems share a common ancestor. The F8 systems of Spirochetes are predicted to be laterally transferred since they group within the Proteobacteria subfamilies rather than forming a separate group. F7 and F8 MCPs are also the only MCPs to have the pentapeptide motif, but F8 MCPs are 34H while F7 MCPs are 36H. CheD is associated with many of the F8 systems by gene neighborhood, but it has not been found to be associated with all of them even with further mirror tree analysis. The F8 CheAs have a P2-II domain. The F8 system of *R. sphaeroides* has been experimentally shown to be necessary for chemotaxis [40]. Although the system is typically transmitted vertically, it also shows a large proportion of lateral transfer instances (Figure 3.2), which suggests it may not be as tightly linked to the flagella as the F7 system.

3.4.8 The F9 System

The F9 system is found in only a few organisms that span distant taxonomic distances including Firmicutes, α and γ-Proteobacteria. F9 CheA proteins have P2-II domains. Its MCPs are uncategorized, but manual analysis of the ones located within the major loci show them to be most similar to 44H MCPs albeit with some small gaps. The MCPs have two MA domains and show no transmembrane regions. The CheW proteins contain three CheW domains instead of the usual single domain. HEAT repeat proteins are also found within the gene neighborhood, and there are no associated accessory proteins. An F9 member has been experimentally studied in *R. centenum* where it was found to control the flagellar biosynthesis [51], but it does not have a TCS-like response regulator like the Alt systems, only a regular CheY protein within its gene neighborhood. A knockout of the F9 system in *V. cholerae* produced no motility defects [57]. It is possible that this system has gained new functions wherever it is present since all

organisms that contain F9 systems also contain at least one other flagellar system more typically found in its taxonomic class.

3.4.9 The F10 System

The F10 system is found in *A. dehalogenans*, *M. xanthus*, and both *Geobacter* species. It is associated with the very long 64H MCPs [34] and the CheAs have P2-I domains. There are no typical accessory proteins associated with F10 systems, but there are HEAT repeat proteins encoded in their chemotaxis loci like the F9 systems. No member of this subfamily has been experimentally characterized.

3.5 Chemotaxis system evolution

3.5.1 Co-evolution of Flagella and Chemotaxis

We propose that the chemotaxis system evolved specifically to regulate the flagellar system, and that both systems were present in the bacterial common ancestor. It has been argued that the complexity of both systems implies that they must have arisen later, and even our initial hypothesis was that the simpler Tfp and Alt chemotaxis systems were the predecessors of the more complex flagellar chemotaxis systems. However, the data supports a reductive scenario, which is in agreement with other claims that evolution can also involve simplification over time of a more complex ancestral system [212,213]. Phylogenetic analysis of FlhA shows evolutionary links between flagella and flagellar chemotaxis system subfamilies. As expected, the FlhA tree is mostly consistent with our 16S tree, given the complexity of the organelle. The FlhA tree shows a subfamily of proteins one of which is a member of the experimentally characterized lateral flagella system of *Vibrio parahaemolyticus*. In this study, the lateral flagella family is found in four members of Enterobacteria, all of which group together. Many members have N-terminal deletions that are predicted to result in non-functional proteins. The lateral flagella FlhA sequences group with the γ-Proteobacteria standard polar flagella, and

lateral flagella have been laterally transferred to two members of β-proteobacteria. A previous study identified a full-length lateral flagella FlhA protein in a different strain of *E. coli* and found that lateral flagella proteins are often encoded in a single locus resulting higher lateral transfer tendencies [214]. Recent experimental work has shown that the same chemotaxis controls both the lateral and polar flagella of *V. parahaemolyticus* [215] and *Vibrio alginolyticus* [216]. In addition, the specific CheY shown to be involved in *V. alginolyticus* is homologous to the F6 CheY that has been shown to exclusively control flagellar rotation in *V. cholerae [217]*. Three of the full length lateral FlhA of our study have F6 chemotaxis systems and are predicted to be analogous to the *V. parahaemolyticus* system.

The flagellar class distribution in comparison to the FlhA tree groupings suggests that there was a duplication of flagellar systems in δ-Proteobacteria that was propagated and specialized among α and β/γ-Proteobacteria along with their associated chemotaxis systems. *R. sphaeroides* breaks this paradigm since experimental work has been shown that its laterally transferred flagella is more essential to motility than its vertically inherited system (Rh.sph1 and Rh.sph2, respectively, in Figure 3.16) [218]; however, the unique nature, thus far, of the *R. sphaeroides* chemotaxis systems supports that such divergences are not typical. The laterally transferred flagella is the sole flagella of *Zymomonas mobilis* based on FlhA analysis, and, like *R. sphaeroides*, it has an F7a system and a divergent second system that does not possess a complete component repertoire based on its gene neighborhood. Although *S. alaskensis* also possesses a single flagellar system of lateral origin, it lacks CheZ and the CheA REC domain in its F5 system, which suggests that the system is being lost and the flagella may not be necessary. *T. maritima* and *Aquifex aeolicus* group together near F1 associated FlhA proteins, but it cannot be established if they were laterally transferred like their chemotaxis systems. The Bacteriodetes member, *S. ruber*, has a FlhA that groups with the Planctomycetes, *Rhodopirellula baltica*, that could be the result of lateral transfer or

long branch attraction (not shown) due to unique natures [219], but the lack of flagella in any other member of Bacteriodetes or Chlorobi and the lateral inheritance of the *S. ruber* chemotaxis system suggests that its flagella is not native.

Enterobacteria and *Chromohalobacter salexigens* lack the polar flagella and F6 systems typically found in γ-proteobacteria. Surprisingly, not only did γ-Proteobacteria *C. salexigens* and Enterobacteria most likely received their chemotaxis systems from a

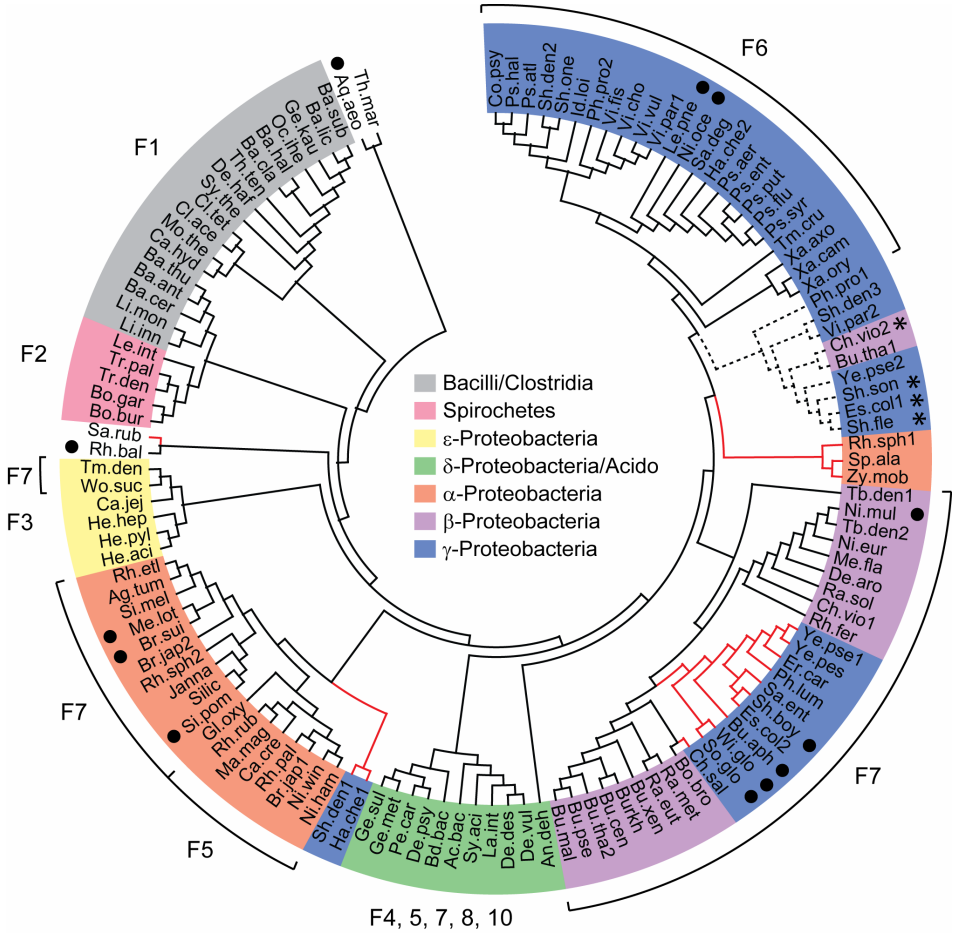


Figure 3.16 Phylogenetic analysis of FlhA shows a pattern of vertical inheritance with discrete lateral transfer events (branches in red). FlhA sequences that are part of the lateral flagella group are identified by dashed branches. Associated flagellar chemotaxis classes are given around the tree. Black circles mark sequences from organisms that lack a flagellar chemotaxis system. Sequences with truncations that are suspected to result in a non-functional protein are marked with asterisks. Sequence identifiers correspond to Table A.14.

member of the β-Proteobacteria genus, *Bordetella*, the FlhA tree supports that the flagella was transferred as well, supporting the tight relationship between the flagella and chemotaxis. Although such a transfer seems unlikely given that flagellar proteins are typically encoded in multiple loci of a genome, a simple MiST query revealed that the flagellar and chemotaxis proteins of *Bordetella brochiseptica* are found in a single locus of its genome which would facilitate a successful transfer of its flagellar and chemotaxis systems to other organisms. We did not perform an in-depth analysis of the flagella, as that is beyond the scope of this study, but a query for all proteins in *B. brochiseptica* and its close relative *Ralstonia eutropha* that contain the description "flagella" in the MiST database [87] showed that only the proteins in the former are found in a single locus. The separation of the flagellar genes into multiple loci in Enterobacteria occurred after the transfer. Maximum likelihood trees of FlhA and CheBR show specific branching of *B. brochiseptica* with *C. salexigens* and Enterobacteria, but they have topology discrepancies elsewhere that are not supported by taxonomic information, which led us to use the trees shown in Figures 3.9 and 3.16.

3.5.2 Adoption of Chemotaxis Components by Tfp and Alt Systems

The Tfp and Alt systems consistently group together in phylogenetic analyses and are more similar to each other than to flagellar systems, implying that they have a common origin. The additional functions of Tfp and lack of tight association with Tfp chemotaxis systems (Figure 3.3) suggest that the ability of chemotaxis systems to regulate Tfp motility was acquired after its original function in the regulation of flagellar motility. Although CheB and CheR are not needed by Tfp chemotaxis systems [50], the presence of CheB and CheR in Alt systems and some γ-Proteobacteria Tfp systems supports the hypothesis that these sequences have been lost by Tfp systems. We predict that the loss of CheB and CheR by Tfp systems is due to the large differences in the mechanisms of flagellar and Tfp based motility. In Tfp motility, the pili attach to the

surface and do not move at the speeds associated with swimming motility [138], which supports that they may be able to employ spatial sensing instead of the temporal sensing required by flagellar motility in liquid [191,220]. This may also explain why two *Helicobacter* species have lost their adaptation systems (and the remaining F3 systems have divergent CheB and CheR sequences shown by their long branch lengths in Figure 3.9) given that organisms with F3 systems live in the gastric muscosa that is very thick and presumably slows movement considerably.

3.5.3 Evolutionary Scenario of Chemotaxis Family Evolution

By combining taxonomic evolutionary information with transition analysis [221] of chemotaxis components and their associated motifs, we have been able to infer a detailed evolutionary history of the chemotaxis system. Transition analysis is simply a way to parse the given data such that the most similar systems are grouped together by way of an inferred evolutionary relationship. Both taxonomic information and transition analysis can be misleading separately, but the combination of the two provides multiple lines of evidence for our evolutionary scenario as well as support for the latest tree of life built by a concatenated alignment of 31 orthologous sequences [211]. The complex F1 systems of Firmicutes likely represent the most ancestral state of the flagellar chemotaxis system, which is supported by multiple evolutionary analyses that place this taxonomic group at the base of the evolutionary tree of bacteria [211,222]. Next the F2 system arose in spirochetes although it fused CheW and CheR, lost CheC and CheD, and had a 4 heptad insertion in the MCPs.

The Tfp system shows vertical evolution within the Cyanobacteria and β/γ-Proteobacteria groups and is present in *D. radiodurans* (not shown in Figure 3.17), whereas the Alt systems are exclusively present in Proteobacteria. The consensus is that Deinococcus and Cyanobacteria predate Proteobacteria [211,221,222]. Thus the Tfp system likely arose out of flagellar systems, and the Alt system came from the Tfp system

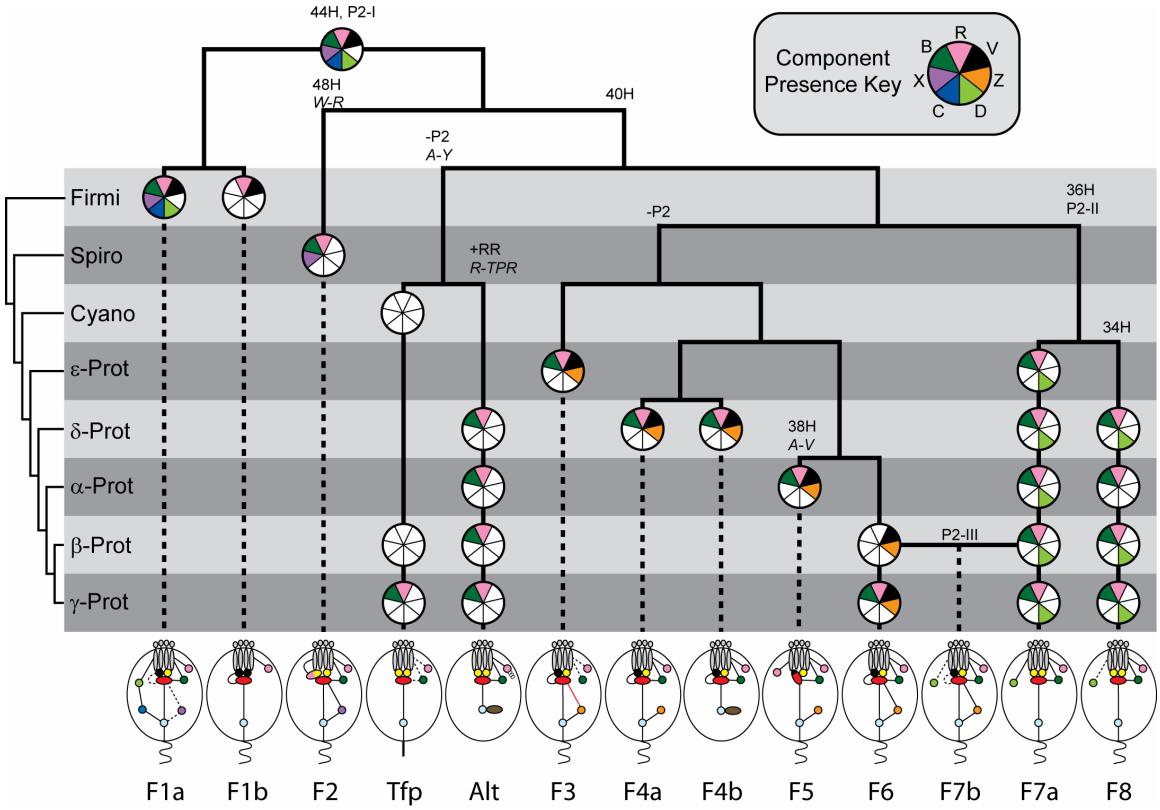


Figure 3.17 A chemotaxis system evolutionary scenario is related to the tree of life. Detailed information about the chemotaxis families and classes allows us to infer their evolutionary history. The tree on the left is a simplified version of the tree of life recently made from a concatenated alignment of 31 universal proteins [211]. Firmi corresponds to Firmicutes. Spiro corresponds to Spirochetes. Cyano corresponds to cyanobacteria. Prot corresponds to Proteobacteria. Circles represent chemotaxis systems present in the lineages listed on the tree to their left. Components present in most members of the systems are color coded in wedges of the circles based on the key at the top right. MCP class, P2 class, and protein fusion information (in italics) is given at nodes of the main tree and assumed to be propogated in all subsequent lineages unless otherwise specified. Cell representations at the bottom represent the chemotaxis classes (F9 and F10 were not included due to conflicting information that does not allow confident placement in the scenario). CheA, CheW, CheY, and MCP proteins are represented in red, yellow, light blue, and gray in the cells, respectively. CheY proteins fused to an additional domain represent response regulators that are more typical of TCSs (Figure 1.1). Accessory components match the color scheme of the component presence key. Essential system interactions are shown in solid lines and interactions present only in some members of a class are shown as dashed lines.

before it lost CheB and CheR. Additionally, there was a 4 heptad deletion in the MCPs of the common ancestor of the Cyanobacteria and Proteobacteria chemotaxis systems before Tfp adopted chemotaxis. Two flagellar systems in the common ancestor of Proteobacteria evolved in the many lineages seen in that group, particularly in δ-Proteobacteria. One lineage gave rise to the F3-F6 systems, and the other gave rise to the F7 and F8 systems. Only the F9 and F10 systems could not be clearly placed in the evolutionary scenario due to a small number of representative sequences as well as unusual motif combinations and/or phyletic distribution.

3.5.4 Speculations on the Origins of the Chemotaxis System

It is accepted that the chemotaxis system shares a common origin with TCSs based upon their shared kinase and receiver domain modules [115]. The general rule of evolution is one of increasing complexity, and just as one component systems arose from TCSs, we propose that chemotaxis systems arose from TCSs based upon their increased complexity. Our proposed evolutionary scenario (Figure 3.17) then goes against the increasing complexity hypothesis since the ancestral systems is inferred to have a larger component repetoire than many of the younger systems. However, there are systems that have decreasing complexity over time [212,213], which is consistent with the notion that fewer system components becomes favorable since it requires less energy expenditure by an organism. The evolutionary scenario also poses that the split between TCSs and chemotaxis systems was an ancient event possibly even in the common ancestor of all bacteria. This goes against the traditional thinking that the highly complex flagellum evolved later in bacteria history, but the vertical evolution of both the flagellum and the chemotaxis system supports the ancient origin of these interrelated systems. A recent study on the evolution of the flagella, reveals that although the flagella contains many components, its complexity is predominantly due to duplications and subsequent

divergence of only a few proteins [223] rather than innovations incorporating multiple unique components.

In the TCS kinases that are associated with divergent CheB and CheR proteins, the coiled-coil region that is predicted to be the site of methylation directly precedes the dimerization domain (Pfam: HisKA and HWE_HK) which has been crystallized in other TCS kinases to reveal a four helical bundle. This architecture configuration makes these proteins attractive targets for the missing link in the evolution from TCSs to chemotaxis systems. We can imagine that this methylated two helix bundle would only need an additional coiled-coil region following the dimerization domain to make the beginnings of an MCP with the remaining kinase and receiver modules of the TCS evolving into CheA and CheY, respectively. The presence of these hybrid systems in organisms that lack chemotaxis systems lends support to this hypothesis, but like classical TCSs, these systems show extensive lateral transfer that obscures their evolutionary history. It is also possible that HKIs adopted CheB and CheR for these new roles after the split between TCSs and chemotaxis systems. For now these systems only represent a functional link, rather than an evolutionary link, between TCSs and chemotaxis systems. Neither evolutionary scenario for the hybrid systems are well supported at this time, but with increasing numbers of sequenced genomes it may be possible to better understand the evolutionary history of the hybrid systems and in turn complete our understanding of the evolution of TCSs and chemotaxis systems.

3.6 Conclusions

Our chemotaxis system analysis supports that there are three chemotaxis system families that originated to regulate different outputs: flagellar motility, Tfp motility, and TCS-like outputs. Paralogy within these families can result in new outputs, such as the two flagellar systems in *R. centenum* only one of which regulates motility, or shared motility regulation such as the multiple flagellar systems of *P. aeruginosa* or *R.*

sphaeroides. Our evolutionary scenario supports that the original function of the chemotaxis system was to regulate flagellar based motility and that the system was born in the common ancestor of bacteria. Tfp and alternative output regulation were adopted later in the evolution of the system, and the flagellar family diverged resulting in 10 distinct classes of flagellar chemotaxis systems. We have determined the primary components and interactions of the Tfp, alternative output, and 10 flagellar classes of chemotaxis systems. We have identified correlations between chemotaxis systems and eight of the 12 MCP length classes. The association of systems with different MCP length classes supports that multiple chemotaxis systems within an organisms are distinctly separate with little cross-talk between systems. Future examination of the sensing repetoire of MCPs could further elucidate the specific environmental signals that are involved in regulating various chemotaxis systems. The interaction predictions should be of use to experimentalists since the interacting components are not always encoded together in the genome and paralogy can obscure the relationships without phylogenetic analysis.

The flagella and flagellar chemotaxis systems show a strong co-evolutionary relationship, which suggests that both originated in the common ancestor of bacteria [223]. The existence of two complex systems in the bacterial common ancestor goes against the traditional thinking that such an organism would have a minimal set of genes with complexity evolving later [224]. However, the roles of gene duplication and gene loss in both systems suggests that a reductive scenario could be an evolutionary paradigm of many more bacterial systems. Indeed, studies support that gene loss plays a major and consistent role in genome evolution that is stronger than gene birth [225,226]. This suggests the last universal common ancestor and presumably the bacterial common ancestor had genomes significantly larger than a minimal set extrapolated from studies of pathogenic bacteria [227].

CHAPTER 4

MOLECULAR EVOLUTION OF CORE CHEMOTAXIS COMPONENTS

Our analysis of the chemotaxis system and its components revealed the diversity and evolution of the system as a whole. Beyond the variety within the core and accessory components of the system that were previously analyzed, there is also a vast sensory repertoire within the MCP family [228], which is primarily determined by various sensing domains that are usually found within the N-teminal region of the MCPs [185,186,206,229,230]. The diversity of the sensing domains is starkly contrasted by the C-terminal signaling subdomain that is the most highly conserved element within the chemotaxis system [34,228,231]. Thus, sensory and signaling modules in MCPs appear to have different evolutionary fates. However, the questions remain: how different are these fates, and what drives the differential domain evolution within an MCP? In order to understand the mechanism underlying its diversity, we sought to quantify the changes at the molecular level by examining the amino acid conservation within the common domains of a highly conserved subfamily of chemotaxis systems that control Tfp motility in cyanobacteria (Tfp-cyano).

We chose to focus on the Tfp-cyano systems because of their high level of conservation. Our large-scale analysis revealed that Tfp chemotaxis systems are found in highly conserved gene neighborhoods and form a monophyletic cluster (Figure 3.2) despite their paralogous relationships. Cyanobacteria are a deep branching phylum of the evolutionary tree [232], and our CheA analysis revealed several cyanobacteria with Tfp chemotaxis systems [156]. At the time of the original study in 2003, there were five complete or draft complete cyanobacterial proteomes with chemotaxis systems available

for analysis: *Synechocystis* sp. PCC 6803, *Nostoc* sp. PCC 7120, *Nostoc punctiforme*, *Thermosynechococcus elongatus,* and *Trichodesmium erythraeum* IMS101 [156]. Since that study, we have restricted our data set to sequences from complete genomes, which includes *Synechocystis* sp. PCC 6803, *Nostoc* sp. PCC 7120, *Thermosynechococcus elongatus,* and the more recently available *Anabaena variabilis* ATCC 29413, *Synechococcus elongatus* PCC 6301, and *Synechococcus* sp. JA-3-3Ab genomes. Tfp chemotaxis systems typically utilize a basic repertoire of chemotaxis components consisting of CheA, CheW, two CheYs, and one MCP, all of which are found in a highly conserved gene neighborhood (Figure 3.7). The conserved nature and apparent vertical evolutionary history of the subset of chemotaxis systems provided an opportunity to use several independent approaches to reveal for the first time the trends in molecular evolution of core chemotaxis components. Although the initial study focused on the evolution of MCP sensing domains, an updated analysis using our current data set reveals new findings that directly reflect the evolution of protein-protein interactions.

4.1 Orthologous Relationships between Cyanobacterial Chemotaxis Operons

In the original analysis, our data set included laterally transferred Alt and flagellar chemotaxis systems from the genomes of *N. punctiforme* and *T. erythraeum*, respectively. Just as the cyanobacterial CheAs predominantly form a monophyletic cluster, trees of similar topology were obtained for the other conserved proteins within their chemotaxis operons that reveal four groups of Tfp-cyano systems. The variety of N-terminal sensing modules between the homologous groups further allowed us to confirm relationships of the sequences within the Tfp-cyano family. The two independent methods, namely phylogenetic analysis of CheA (Figure 4.1) and detailed domain architectures of MCPs (Figure 4.2), provided identical results for the identification of orthologous systems.

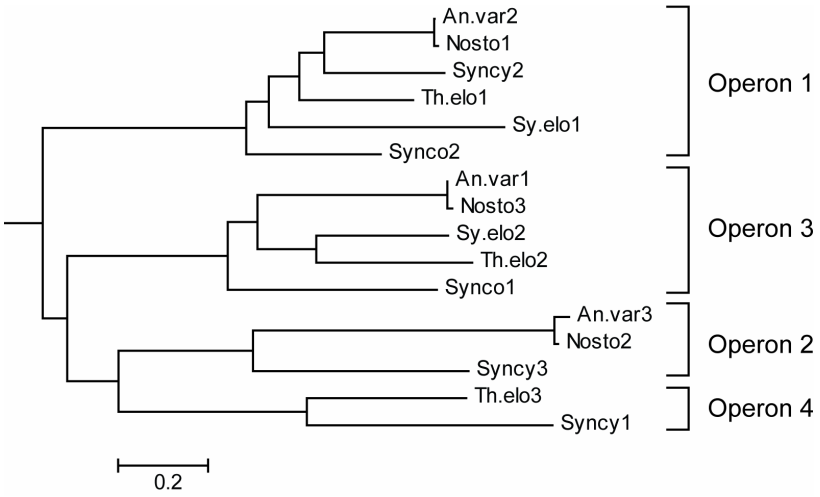


Figure 4.1 The subtree of the cyanobacteria CheA sequences from Figure 3.2 shows the same four groups described in the original study [156]. Sequence identifiers correspond to information in Table A.2.

On the basis of these results and taking into account known trends in evolution of operons [233], a simplified evolutionary scenario supports the existence of four paralogous chemotaxis operons in the cyanobacterial common ancestor, which encoded CheY1, CheY2, CheW, MCP, and CheA (Figure 4.2). CheY2 proteins of Tfp-cyano systems are stand-alone REC domains like all previously characterized CheY, but CheY1 proteins are homologous to the PatA protein involved in heterocyst formation [208] and chemotaxis [63] and contain an additional unique N-terminal domain [207]. Phylogenetic analysis of these unusual proteins confirmed that they are closely related to CheY and are predicted to have a similar function. The Operon 1 *cheA* of *Synechocytis* sp. is split into two individual genes that are located on the chromosome distantly from each other and from the remainder of the operon (Figure 4.2). Nonetheless, both genes have been experimentally proven to be necessary for chemotaxis [63,234]. Consistent with the original analysis, Operon 1 appears to be the most conserved and is present in all species (Figure 4.1). Operon 2 is present in, *A. variabilis*, *Nostoc* sp., and *Synechocystis* sp. of the current data set, but was also found in draft genomes of *N. punctiforme* and *T.*

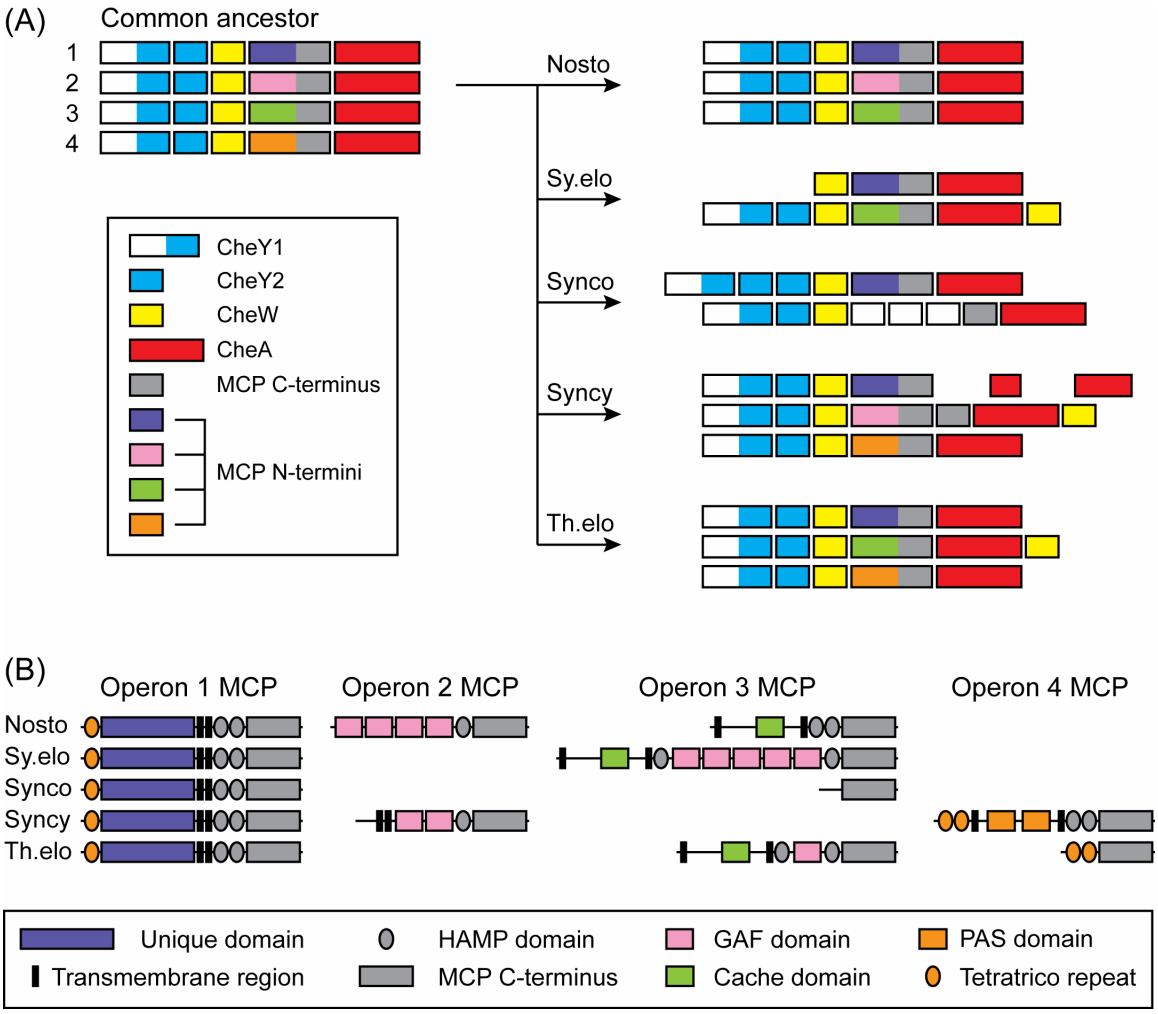


Figure 4.2 (A) Classification of operons (1,2,3,4) is based on the results of phylogenetic analyses of the CheA (Figure 4.1), CheW, CheY, and MCP signaling domains of the proteins encoded in the operons, and on the presence of specific N-terminal modules in MCPs encoded within the operons. Nosto, Sy.elo, Synco, Syncy, and Th.elo correspond to *Nostoc* sp. PCC 7120, *S. elongatus* PCC 6301, *Synechococcus* sp., *Synechocystis* sp., and *T. elongatus* genomes, respectively. The white region of CheY1 indicates its unique N-terminal domain. Other white boxes indicate genes not associated with chemotaxis due to a transposon disruption of Synco's Operon 3. *A. variabilis* data is not shown because it is nearly identical to that of *Nostoc* sp. as shown in Figure 4.1. Gene neighborhoods are not drawn to scale. **(B)** Domain architecture of the MCPs associated with each operon family (not to scale). The numbering and color code are the same as in (A).

erythraeum in the original study [156]. Operon 2 was the most prevalent Tfp-cyano system other than Operon 1 in the original study, but the new study shows Operon 3 to be more prevalent. Operons 2 and 3 may have similar functions based on domain architecture analysis that will be discussed later, which could account for the discrepancies between the data sets. Operon 3 is present in all species of the new data set except *Synechocystis* sp., but the original study also showed it was absent in *T. erythraeum*. As seen in our original study, Operon 4 is present only in *T. elongatus* and *Synechocystis* sp.

4.2 Domain Birth, Death and Innovation in MCP Sensory Modules

Domain architecture analysis revealed that MCPs from paralogous operons have very different N-terminal sensory modules, whereas sensory modules in MCPs from orthologous operons are quite similar (Figure 4.2). The N-terminal module of the orthologous MCPs from Operon 1 (shown in dark blue) is unique (i.e. no homologous domains were detected in any publicly available database) except for a small N-terminal region that shows homology to a tetratricopeptide repeat domain (a known site for protein–protein interactions [182]) in domain architecture analysis and PSI-BLAST searches. Fold recognition using the 3D-PSSM program [235] and secondary structure prediction using the JPRED2 program [109] suggest that the rest of this module comprises a globular and mostly α-helical domain rather than an unstructured region. The N-terminal modules of the orthologous MCPs from Operon 2 consist of several GAF domains, known phototransducing elements [188]. Indeed, the Operon 2 MCP from *Synechocystis* sp. was predicted [229] and experimentally shown [63,236] to act as a receptor for phototaxis. All of the N-terminal modules of the MCPs from Operons 3 contain a Cache domain [206] with the exception of that from *Synechococcus* sp. due to a transposon insertion (Figure 4.2). Some Operon 3 MCPs also contain cytoplasmic GAF

domains as seen in Operon 2. Operon 4 MCPs have tetratricopeptide repeat domains that were identified in PSI-BLAST searches, and one member also contains two periplasmic PAS domains (also identified by PSI-BLAST). Although there are common domains within each group of MCP sensory modules, dramatic consequences of domain shuffling, birth, and death in the sensory modules of MCPs from Operons 2, 3 and 4 are apparent, whereas no such changes can be seen in MCPs from Operon 1 (Figure 4.2).

4.3 Evolutionary Rates of Chemotaxis Modules

Our analysis shows that Operon 1 is the most highly conserved given its presence in all five cyanobacteria, and its importance is further confirmed by studies in *Synechocystis* sp., which showed that mutations in the CheA or MCP of its Operon 1 abolish its motility [63,234]. The high conservation of Operon 1 makes it the best choice to examine the rate of evolution of orthologous components. No obvious changes (e.g. domain shuffling, large insertions and deletions) were observed in the domain organization of the sensory modules of Operon 1 MCPs. Therefore, we have measured the percentage of sequence identity in these modules and compared it to that in other conserved modules of the chemotaxis proteins. Pairwise global alignments were performed between all orthologous proteins or modules and the resulting sequence identities were averaged. The Operon 1 of *Synechococcus* sp. has a duplication of CheY2 protein (Figure 4.2), but phylogenetic analysis identified the CheY2 protein closest in proximity to CheW as the one most closely related to the other Operon 1 CheY2 sequences for use in our sequence identity analysis. Given the nearly 100% identity between most of the Tfp-cyano orthologs in *A. variabilis* and *Nostoc* sp. (Figure 4.1), the former were excluded from the latest analysis.

The original study found that the conserved components of Operon 1 had an average sequence identity above 60%, but the MCP N-terminal modules were strikingly different with a sequence identity below 25% [156]. Consistent with that data, our

updated analysis of Operon 1 shows that the majority of the conserved domains have an average sequence identity above 40%, and MCP N-terminal modules have less that 20% average sequence identity (Figure 4.3). A similar trend was observed for other orthologous operons in the original study, but the significant diversity of the MCP N-terminal modules made it impossible to analyze any other operons in our updated data set. The inability to find sequence similarity for the other operons further supports the notion that in many, if not most, MCP orthologs the N-termini diverge beyond the detectable level of sequence similarity. We can therefore conclude that the N-terminal modules, which are responsible for sensory properties of MCPs, have significantly accelerated evolutionary rates.

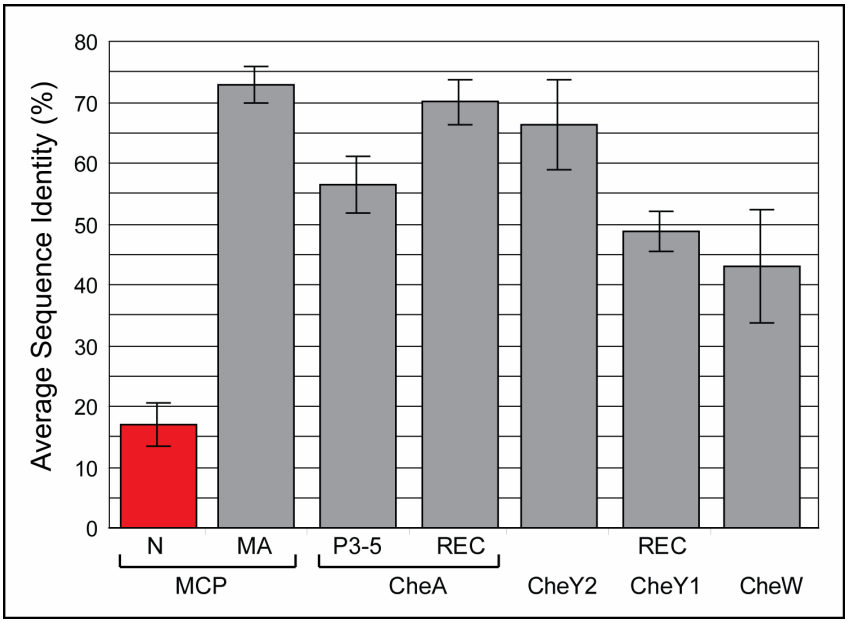


Figure 4.3 Average sequence identity of conserved modules from Operon 1 components. MCP-N is the sensory module. MCP-MA is the signaling module. CheW is the entire CheW protein. CheA-P3-5 is the dimerization, ATPase, and CheW domains of CheA (Figure 3.1). CheA-REC is the C-terminal REC domain of CheA. CheY2 is the entire CheY2 protein. CheY1-REC is the REC domain of CheY1. CheY1 and CheY2 averages do not include sequence data from *S. elongatus* since both proteins are absent from Operon 1 (Figure 4.2).

4.4 Biological Implications

Proteins or their domains that are involved in protein–protein interactions show a shared rate of evolution. On the other hand, proteins that do not interact with each other have different evolutionary rates even if they are encoded in the same operon and participate in the same regulatory pathway [237]. The sensory modules of MCPs do not participate in crucial protein–protein interactions within the chemotaxis pathway, as do their signaling modules and all other proteins encoded in chemotaxis operons [16,238]. The absence of physical constraints imposed by such interactions account for the drastic increase in the evolutionary rate of amino acid change and domain shuffling in these sensory regions. The evolution of MCPs follows a domain birth, death and innovation model (BDIM) applied to a multidomain network [239], in which the conserved C-terminal module serves as a hub connecting the receptor to other proteins in the signal transduction pathway. Rapid BDIM-type evolution of the sensing modules of MCPs creates a greater diversity of sensing capabilities and aids the organism's ability to respond to a wider variety of extracellular and intracellular signals. It is likely that this trend will be observed for other sensory receptors in prokaryotes.

Analysis of the updated data set provides additional evidence for the role of protein-protein interactions in maintaining sequence conservation. While the majority of chemotaxis modules other than the MCP N-terminal module were thought to be well conserved, the Hpt and P2-Hpt domains did not show the expected amounts of sequence conservation and both had very large standard deviations. The original study showed the Hpt domain to be of a similar conservation level as the dimerization, ATPase, CheW, and REC domains of CheA. The P2-Hpt domain was discovered after the original study and is predicted to be a CheY binding site that originated from a Hpt domain duplication. Although the P2-Hpt domain was not expected to have high sequence similarity analogous to the P2 domains of flagellar chemotaxis systems, the standard deviation of the P2-HPT average identity was similar to that of the MCP N-terminal module. A closer

inspection of the data set showed that the low sequence similarity and high standard deviation are the result of the inclusion of Hpt and P2-Hpt domains from *S. elongatus*, the same organism that lacks CheY1 and CheY2 in Operon 1. When these sequences were removed from the analysis, the average sequence identity increased and the standard deviation was similar to those of the other chemotaxis modules. CheY is known to interact with the CheA Hpt domain [32], so it is not surprising that the loss of both CheY proteins has resulted in sequence divergence of the Hpt domain. The similar pattern seen in the analysis of P2-Hpt supports our prediction that it is a CheY binding domain. The high conservation of the remaining Operon 1 modules of *S. elongatus* suggests that the loss of CheY1 and CheY2 was a recent event. We predict the remaining modules will also lose sequence conservation, and given the importance of Operon 1 based on

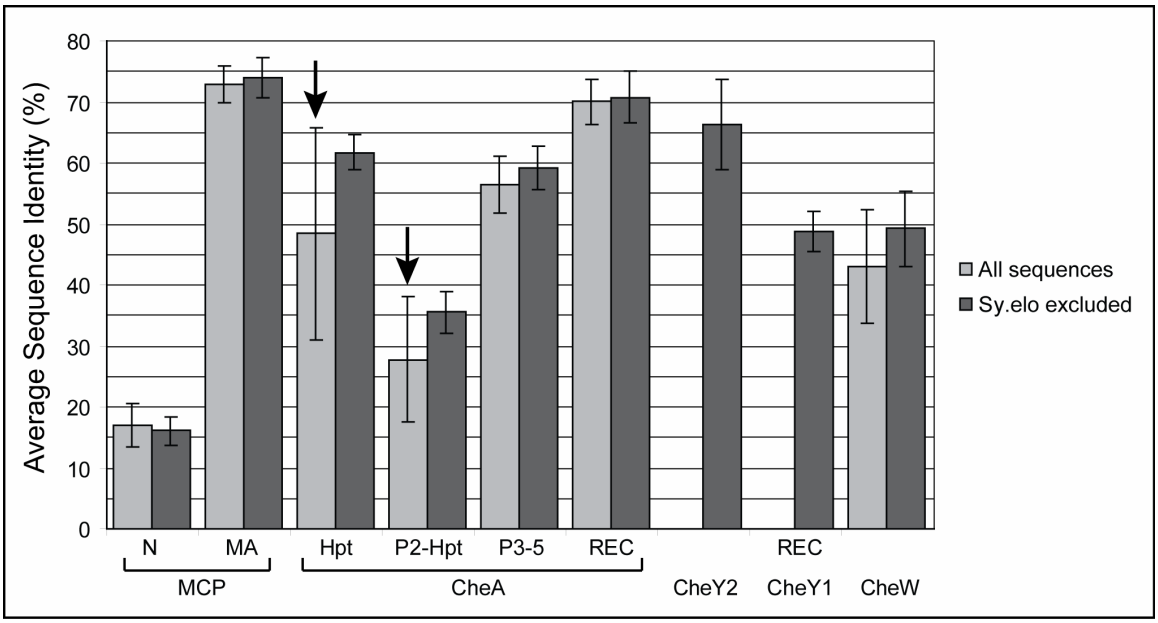


Figure 4.4 A comparison of the average sequence identities of chemotaxis Operon 1 modules from *Nostoc* sp., *S. elongatus*, *Synechoccus* sp., *Synechocystis* sp., and *T. elongatus,* with the data set excluding *S. elongatus*, shows rapid divergence of the Hpt and P2-Hpt domains that is correlated to the absence of CheY1 and CheY2 in the latter organism. Arrows highlight the low sequence identity and high standard deviation in the columns of interest.

conservation within chemotactic cyanobacteria and experimental studies, it is also likely that the remaining chemotaxis system (Operon 3) could also be lost.

4.5 Conclusions

Comparative genomic analysis revealed that sensory domains of cyanobacterial chemoreceptors evolve much faster than their signaling domains and the rest of the chemotaxis signal transduction pathway. Fast sequence evolution in sensory domains could lead to new sensing capabilities of chemoreceptors. It is likely that this trend will be observed for other sensory receptors in prokaryotes. The rapid sequence evolution of sensory domains can be attributed to their limited involvement in protein-protein interactions. We observed that the recent loss of CheY proteins in a chemotaxis system resulted in the rapid divergence of its cognate interaction domain in CheA, the Hpt domain, in addition to rapid divergence of a novel domain P2-Hpt that we predict binds CheY. These results support that interactions play a significant role in shaping the rates of protein evolution at the sequence level.

CHAPTER 5

CONTACT SITE PREDICTION

Protein-protein interactions are essential for mediating information transfer in complex systems like signal transduction and metabolic pathways. Comparative genomics methods such as phylogenetic profiles, gene neighborhood, gene fusion, and mirror tree analysis can all be used to predict protein-protein interactions [78]. The topology of a protein interaction network provides general information about how the network functions, but to understand the mechanisms behind its functioning we need to gain a better understanding of how the proteins are interacting at the molecular level. Co-crystals have been invaluable in understanding the mechanisms of protein-protein interactions, but they typically capture a single rigid state of an interaction, which may not reflect all of the residues important to an interaction *in vivo*. Structural and experimental data about protein interaction surfaces in conjunction with genomic data provide an opportunity to develop new predictive methods to identify potential protein-protein contact sites and enable design-driven studies of interaction characteristics.

Contact site prediction methods currently fall into two main categories: methods that predict contact sites of a given structure [86,240,241] and those that predict contact sites of a given sequence [85,242-247]. Currently structural prediction methods perform better than their sequence-based counterparts, but sequence-based methods have a broader impact since there is currently far more sequence data than structural data for most proteins. There are two classes of sequence based prediction methods: learning methods that are first trained on co-crystal information before they predict interaction sites from sequence [246,247], and methods that exclusively use sequence data without structural information [85,242-245]. The high-throughput capabilities of learning methods are undoubtedly of great benefit to scientists who wish to understand the general

properties of protein-protein interactions, but these methods often lack the high specificity needed for efficient experimental analysis. The flexible and semi-manual nature of sequence analysis without structural data provides an opportunity to develop contact site prediction methods that are easy to implement and immediately useful for experimental work.

Exclusively sequence-based contact site prediction methods include proline bracket analysis [244], hydrophobicity distribution [243], correlated mutation analysis [85,86], and functional residue identification [242], but the latter two are the most extensively studied and relevant to this study. Protein-protein interactions are maintained over the course of evolution because the amino acid residues involved in an interaction are more resistant to change than other surface residues. This observation and experimental second site suppressor studies have been the foundation for computational approaches to identify correlated mutations within and between families of proteins, which could be the sites of intraprotein [83] and interprotein [85,86] interactions, respectively. Furthermore, the concatenated or virtually concatenated alignments of correlated mutation methods exclude paralogs from analysis since the method requires a one-to-one relationship [86]. We can easily include all proteins of interest by including phylogenetic analysis similar to the mirror tree method of protein interaction prediction [74,75].

Our goal is to further develop a contact site prediction method from an idea that was originally studied in Casari et. al. [242] where they proposed that residues exclusively conserved within subfamilies of proteins family served functions specific to those subfamilies, which could include binding other proteins or cofactors although they did not analyze the interacting partners. Interprotein contact site prediction methods typically involve the identification of contact sites between core proteins that are always present in a given interaction network. Many types of protein interaction networks have accessory components that are present only in some members of the network family. In

comparing homologous interaction networks, we predict core components that interact with accessory components have conserved amino acid residues that maintain the additional interactions, which are not conserved in the core components that lack the accessory interactions (Figure 5.1). The additional conserved residues within the subset of core components that interact with a given accessory component are predicted to be reflected in the topology of the phylogenetic tree of the core component. Residues exclusively conserved within a subfamily associated with an accessory interaction are predicted to interact with the accessory protein. All conserved residues of an accessory component are predicted to be involved in interactions with core components in addition to the structure and function of the protein.

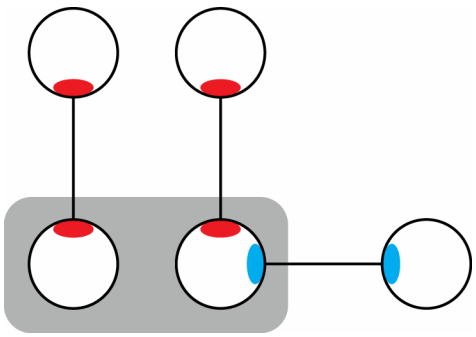


Figure 5.1 The subfamily subtraction methodology for identifying protein-protein contact sites in core-accessory interactions. The circles represent proteins. Red spots represent contact sites in core-core interactions. The contact sites associated with a core-accessory interaction present in a subset of systems are shown in blue. By comparing core proteins that have the accessory protein with those that do not, we aim to identify accessory interaction residues.

5.1 Subfamily Subtraction Method for CheY-CheZ Interaction

All CheZ proteins are predicted to interact with CheY. Consensus analysis was carried out on the full length CheZ alignment after sequences that lack cognate CheY partners were removed. Conserved positions were identified as those that are conserved

at 95% or higher since CheZ is not a highly conserved protein as a whole. Since only four residues of the 214 a.a. *E. coli* CheZ protein were predicted to have 0% solvent accessibility, positions predicted to have 5% solvent accessibility or less were considered to be buried in our analysis. Conserved positions that are predicted to be solvent accessible are also predicted to be CheY contact sites.

Many CheY proteins are found in genomes that lack CheZ [15,33]. An alignment of the flagellar CheY proteins and subsequent multiple alignment revealed a subfamily of 71 sequences (CheYz subfamily) in the CheY phylogenetic tree that are predicted to interact with CheZ. 54 of them are encoded immediately adjacent to *cheZ*, and 13 of the remaining 15 sequences have at least one CheZ encoded elsewhere in their genomes. Tfp CheY proteins (Chapter 3.3.6) were excluded from the analysis since they work with an entirely different motility system. We analyzed the consensus sequence for the full length alignment of the isolated CheYz subfamily minus the four sequences that lack a cognate CheZ in their genomes. Consensus was also determined for the full length alignment of the remaining 221 CheY sequences (CheYz background) after two sequences were removed due to deletions in the N- or C-terminal of the core portion. Positions that are conserved at 98% or higher in the CheYz subfamily were compared to the conservation level of the same physicochemical properties of the same positions in the CheYz background. The 98% threshold was chosen to account for any minor differences that might occur in some sequences. The 16 residues predicted to have 0% solvent accessibility in the 129 a.a. CheY protein of *E. coli* were classified as buried for our analysis. The background conservation level of the conserved CheYz positions ranges from 7-100%. When the hydrophobicity and burial prediction was included in the background conservation analysis, we found that the residues that are universally conserved are hydrophilic in nature and predicted to be on the surface consistent with their probable role in catalysis or structural residues like glycines and prolines. The next most highly conserved CheYz positions are buried and/or very hydrophobic, which is

typical of hydrophobic core residues. The conserved CheYz residues with the lowest level of background conservation are of mixed hydrophobicity and usually not buried, consistent with the nature of non-oligomeric protein-protein interfaces. Looking at a chart of the background conservation in addition to hydrophobicity and burial analysis, we chose the conserved CheYz residues that have a background conservation of 59% or less and are not buried as our predicted CheZ contact sites in order to maximize the amount of hydrophilic residues that we expect to be a significant portion of the interface. The analysis was repeated after removing the subfamily involved in CheC and CheX interactions from the background analysis. Even though the new background exclusively

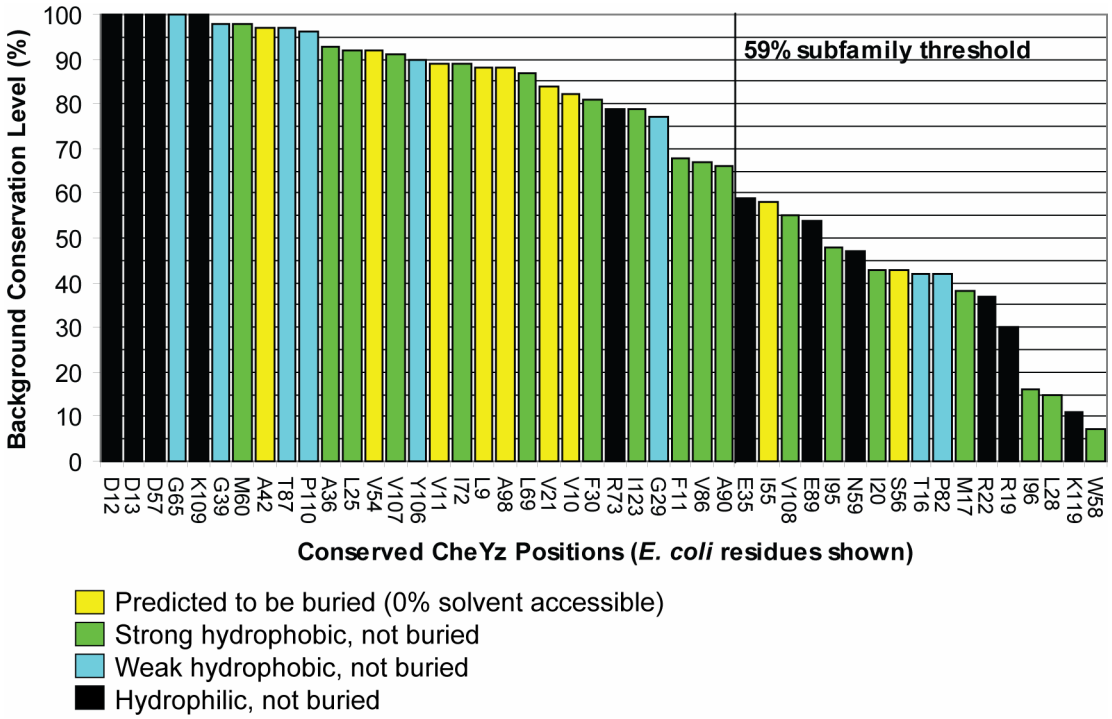


Figure 5.2 A graph of the conserved residues within the CheYz subfamily in comparison to their conservation levels in the remaining family members. Hydrophobicities greater than zero, between 0 and -2, and less than -2 are considered strong hydrophobic, weak hydrophobic, and hydrophillic, respectively [248]. As predicted, we see enzymatic and structural residues (hydrophillic and glycines/prolines, respectively) as the most highly conserved, followed by the hydrophobic core residues, and last the predicted contact sites. Protein-protein interfaces usually contain mixed hydrophobic and hydrophillic residues in heterocomplexes [249].

consisted of Proteobacterial sequences (like the CheYz subfamily), the subfamily threshold level was almost identical as before (60%), but there was a minor effect on the specificity and sensitivity of the method (Chapter 5.2.2).

5.2 CheY-CheZ Contact Site Prediction

We predicted the contact sites between a core protein, CheY, and accessory protein, CheZ, of the prokaryotic chemotaxis signal transduction system because the interaction has been well characterized structurally [25,30] and experimentally [201,202,250-253]. The single domain CheY response regulator is the primary output protein of the prokaryotic chemotaxis system, which interacts with the motor system. Phosphorylation of CheY increases its affinity for the motor system it regulates, and CheZ is a phosphatase of CheY-P [153]. While all chemotaxis systems have a response regulator, CheZ had only been identified in some β/γ-Proteobacterial chemotaxis systems until a recent identification of a divergent CheZ protein in the ε-Proteobacteria, *Helicobacter pylori* [197]. Our recent study of the evolution of the chemotaxis system revealed new CheZ sequences in some members of α and δ-Proteobacteria (Figure 3.11), and a comparison of CheY and CheZ trees shows that the two proteins share the same evolutionary history (Figure 5.3). The CheY/CheZ co-crystal structures [25,30] are used to validate our predictions, and experimental studies in *E. coli* further show the greater importance of the conserved contact sites than unconserved contact sites and the relevance of neighboring conserved residues that play indirect roles in maintaining the interaction.

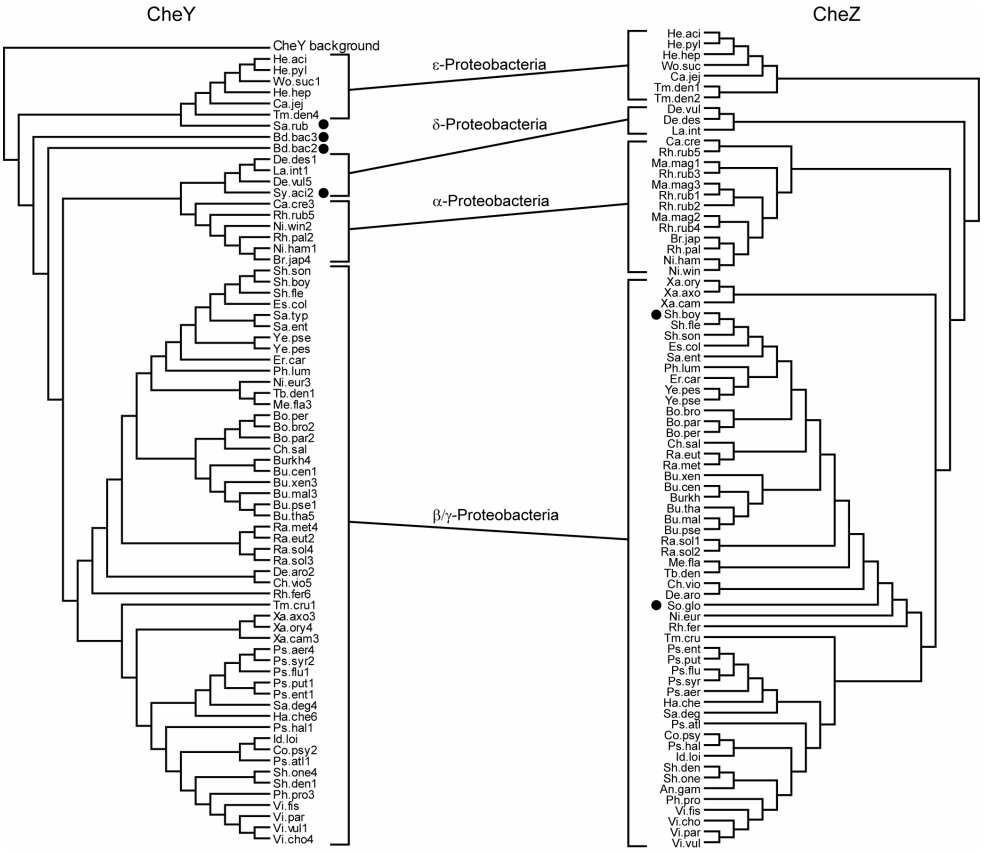


Figure 5.3 A dendrogram of the CheYz subfamily is shown with the outlying 224 CheY sequences shown as a collapsed branch. The CheYz subtree and CheZ tree topologies have shared subfamilies that contain the same species for each Proteobacteria class. Sequences marked with a black circle lack an interaction partner and were excluded from their respective alignments for conservation analysis. CheY and CheZ identifiers correspond to Tables A.11 and A.6, respectively.

5.1.1 Prediction of CheY contact sites on CheZ

Sequence analysis predicted eight CheY contact sites in the CheZ sequence, and there is a partial overlap between these residues and the 17 residues collectively identified as CheY contact sites in three different co-crystals. The structural study identified 11 contact sites that were not found to be conserved in our sequence analysis. The lack of pressure on these residues to resist mutation suggests that they may not be as important as the conserved residues in maintaining the interaction and have potentially undergone covariation with their CheY contact residues. The CheZ structure reveals that

it exists as a coiled-coil dimer and six of the predicted contact sites are on the surface, five of which were previously confirmed as contact sites by structural analysis [25,30]. One predicted CheY contact site, F141, is instead part of the dimerization interface as are five of the seven conserved CheZ residues that were predicted to be buried. Of the two conserved residues that were falsely predicted to be buried, one was previously confirmed to be a contact site in structural studies, L208, and the other, L144, is adjacent to confirmed contact site residues even though it was not identified as a contact site in previous structural studies. We identified that L144 is involved in hydrophobic interactions with the CheYK109 sidechain, given the 3.84 Å between L144 and the CheYK109 beta carbon. A conserved glycine, G188, that was predicted to be a contact site resides in an unresolved loop of the CheZ structure, and it is most likely involved in a functional critical turn rather than CheY contact.

A conserved glutamate, E74, was exclusively identified as a contact site by our sequence analysis, but it lies adjacent to two residues, E67 and N71, that were exclusively identified as contact sites by structural analysis. Although E74 is 4.8 Å away from the salt bridge partner of E67 and N71, CheYR19, *in vitro* in the CheY-CheZ co-crystal, the complete conservation of a negative residue in the E74 position in comparison to the poor conservation of the E67 and N71 positions led us to predict that E74 might be the more important contact *in vivo*. Mutations of each residue (E67R, N71R, and E74R) were studied by swarm plate analysis. The N71R mutant has over 80% of the wild type (wt) swarm rate and is still chemotactic. Both glutamate mutants have significantly reduced swarm rates and are not chemotactic, but the E74R mutant is almost non-motile with less than 10% of the wt swarm rate while the E67R retained ~40%-50% of the wt swarm rate. These experimental results support that the conserved surface residue E74 plays a more important role in chemotaxis than the structurally identified contact sites adjacent to it. Given the location of E74, this difference most likely can be attributed to its predicted role in CheY interaction.

5.1.2 Prediction of CheZ contact sites on CheY

Our subfamily subtraction method predicted 15 CheZ contact sites in the CheY sequences that partially overlap with the set of 16 contact sites identified by previous structural studies. Nine of the predicted contact sites were confirmed by those studies, but all of the remaining six are found on the surface of CheY. Four of the seven contact sites exclusively identified by structural studies are not specific to CheZ interactions as our analysis identified them as highly conserved among all CheY proteins, not just those that interact with CheZ. The remaining three structurally identified contact sites are not highly conserved within the CheYz subfamily; thus, they are predicted to play a lesser role in maintaining the CheY-CheZ interaction or to have undergone covariant changes with their interacting partners on CheZ.

Of the six predicted contact sites exclusively identified by sequence analysis, four (R22, E35, W58, and N59) are directly adjacent to confirmed contact sites on the surface. Mutations of residues W58 and N59 result in a loss of chemotaxis and significant swarm defects, and they are predicted to play a role in maintaining the proper activation and conformation for CheY dephosphorylation given their close proximity to the phosphorylated D57 and the critical CheY activating residue Y106 [254]. Residues R22 and E35 play a peripheral role in maintaining the CheY/CheZ interaction, by stabilizing an alpha helix that interacts with CheZ at the catalytic interface via a salt bridge formed between the two residues. Mutations in R22 and E35 do not destroy chemotaxis, but they do reduce swarm rates by 50% or more than that of wild type cells. The remaining two residues exclusively identified by sequence analysis, L28 and P82, are not predicted to play a role in CheZ interaction. P82 is far from either CheZ interaction surface and is involved in a critical turn that is generally conserved in all CheY proteins, but its location can vary by a few positions in the sequence, which kept it from having a high background conservation. L28 is not directly adjacent to any contact sites, but it is near the catalytic

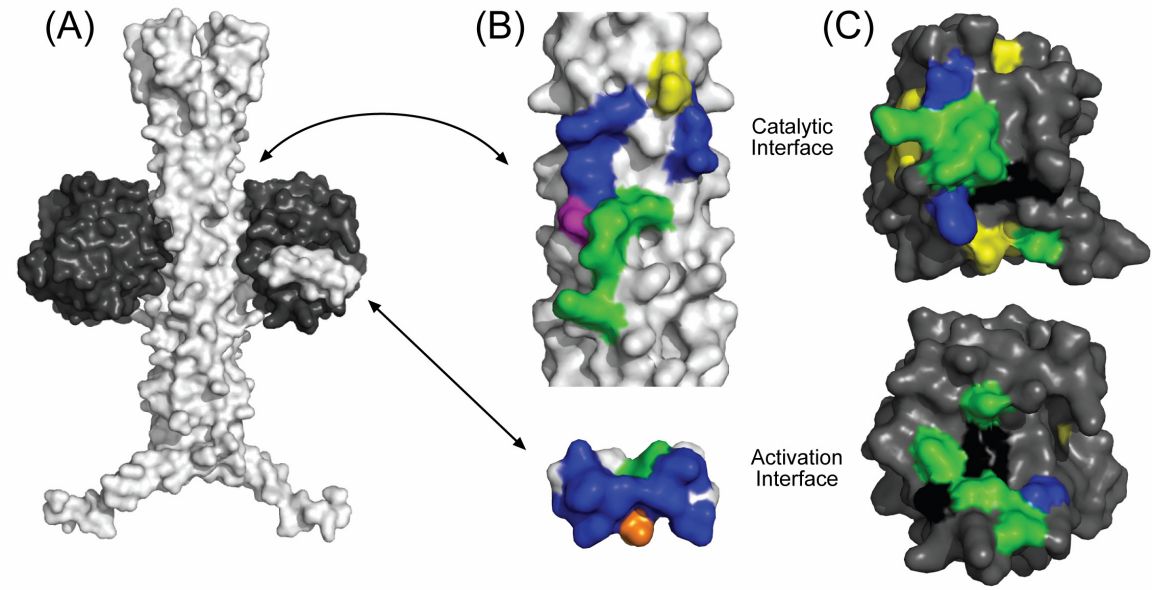


Figure 5.4 (A) Surface view of the CheZ dimer (white) and a two bound CheY (black). **(B)** Close up views of the two surfaces of CheZ that contact CheY. Contact sites exclusively identified by structural analysis, E67, N71, M136, M137, D140, A200, Q202, V205, L209, L212, and F214 are blue. The contact site exclusively identified by sequence analysis, E74, is yellow. Contact sites identified by both sequence and structural analysis, Q142, D143, Q147, R151, D206, are green. Conserved residue L208 (shown in orange) was previously identified as a contact site in structural studies although it was falsely predicted to be buried in our sequence analysis. Conserved residue L144 (shown in magenta) was also falsely predicted to be buried, and we have found that it is a true contact site even though it was not identified as one in previous structural studies. **(C)** Close up views of the two surfaces of CheY that contact CheZ. Contact sites exclusively identified by structural analysis, F14, N23, and A99, are in blue. Contact sites exclusively identified by sequence analysis, R22, L28, E35, W58, N59, P82, are in yellow. Contact sites identified by both sequence and structural analysis, T16, M17, R19, I20, E89, I95, I96, V108, and K119, are green. Residues that were identified as contact sites in structural studies, but found to be highly conserved among all CheY proteins in our analysis, D12, A90, Y106, and K109, are shown in black.

interface. Mutations of L28 result in no significant chemotaxis or swarm defects, and the reason for its conservation in this subfamily is not known.

When the CheY subfamily associated with F1 systems and CheC interaction was removed from the background data set used for CheYz analysis (so that primarily Proteobacterial CheY sequences were being compared), only three predicted contact sites were lost in comparison to the full background set. One lost prediction was predicted to

be a contact site in the co-crystal, and one was predicted to be a contact site exclusively based upon sequence analysis. The third lost prediction was the P82 structural residue. The increased specificity suggests that this method does not require large evolutionary distances. The lost contact sites may play a more general role in CheA-CheY interaction in addition to CheY-CheZ interaction since the lost residues were from the main interaction face only, which is also predicted to be the CheA-CheY interface.

Table 5.1 A comparison of the specificities and sensitivities of the subfamily subtraction method in identifying CheY-CheZ contact sites. Specificity is defined as the number of true predicted contact sites divided by the sum of the true predicted contact sites and the false predicted contact sites. Sensitivity is defined as the number of true predicted contact sites divided by the sum of the true predicted contact sites and the true contact sites that were not predicted. Predictions validated by experimental and structural data show higher sensitivities and specificities than those validated by structural data only. The CheZ contact sites predicted on CheY show a higher specificity, but lower sensitivity, when the proteobacterial background (Yz-prot) was used instead of the full background (Yz-all).

Validation	**Structure + Experiment**			**Structure Only**		
Protein	**Yz-prot**	**Yz-all**	**Z**	**Yz-prot**	**Yz-all**	**Z**
Specificity	0.909	0.867	0.750	0.636	0.600	0.625
Sensitivity	0.526	0.650	0.333	0.438	0.563	0.294

5.3 Subfamily Subtraction Follow Up Analyses

We chose to apply the subfamily subtraction method that we developed in the CheY-CheZ analysis to other core-accessory interactions of the chemotaxis system. CheY binds CheA at the Hpt domain in a core-core interaction since the Hpt is an essential part of all CheA proteins, but the CheY interaction with the P2 domain is not conserved among all chemotaxis systems (Chapter 3.2.1). There are three available co-crystals for the validation of CheY-P2 predictions [20,21,26]. The CheC and CheD chemotaxis proteins are experimentally known to interact together and have been co-crystallized [29], and there are subsets of each family member that do not interact with each other [194] (Chapters 3.3.2 and 3.3.4). Although CheY and CheC have not been co-

crystallized, there are individual crystals of each and experimental information about their active sites allows us to predict the main faces of each protein involved in the interaction

5.2.1 CheY-P2 Contact Site Analysis

Of the three P2 domain classes, two have been co-crystallized with their cognate CheYs: P2-III from *E. coli* [20,21]and P2-I from *Thermotoga maritima* [26]. The co-crystal from *T. maritima* revealed that the interaction mechanism between its P2 and CheY is very different from the P2-CheY interactions of *E. coli* [26]. Not only is the interaction orientation of the two crystals different, but as also shown in sequence analysis, the P2 domain of *E. coli* is significantly reduced in size due to the deletion of an alpha helix (Figure 3.6). There is also very little sequence conservation across P2 domains, even within subfamilies. Given that the P2 domain has only been shown to interact with CheY, all of its conserved residues are predicted to interact with CheY.

The CheY-P2-III interaction typified by the *E. coli* chemotaxis system is present only in the 28 members of our F7b data set. Previous analysis showed that the P2-III domain is tightly associated with the F7b chemotaxis family, possibly due to steric constraints of the CheA-CheZ interaction (Chapter 3.4.6). 27 of the 28 F7b members have an associated CheZ protein; *Burkholderia thailandensis* has two F7b systems but only one has retained CheZ. Sequence analysis of the 27 P2-III domains shows only two positions conserved at 100%. Of the eight positions conserved in at least 26 of the 27 members, only two are confirmed interaction sites. A conserved aspartate, which is found adjacent to a non-conserved aspartate identified as a contact site in the co-crystal structure, is predicted to play a role in the interaction, like the conserved glutamate of CheZ identified in the CheY-CheZ analysis. The remaining five conserved residues are aliphatic residues involved in core folding interactions that were not predicted to be 0% solvent accessible. Most of these core residues were identified as 5% solvent accessible,

but use of that threshold loses one of the conserved contact sites (F14) in the crystal structure. The high conservation level of the closely related CheYs predicted to bind P2-III makes it impossible to discriminate conserved CheY residues specific to P2-III interaction.

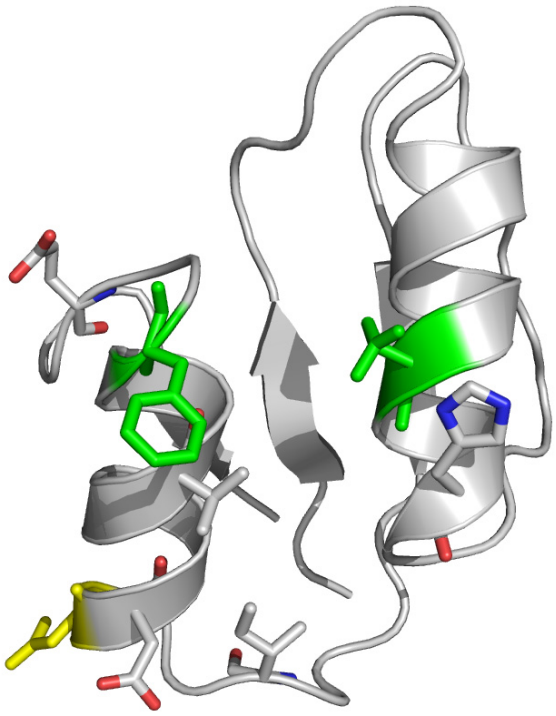


Figure 5.5 A cartoon representation of the P2-III domain of *E. coli* [20] with a stick representation of the amino acids that interact with CheY based on co-crystal data. The yellow residue is a conserved aspartate exclusively identified by sequence analysis. The green residues were identified in sequence and co-crystal analyses. The remaining residues were only identified by co-crystal analysis

Analysis of the CheY-P2-I reveals the opposite problem from the CheY-P2-II analysis. The P2-I domain is in many diverse taxa including Firmicutes, Spirochetes, Proteobacteria, and Archaea of the F1, F2, and F10 chemotaxis subfamilies. Consensus analysis shows only one highly conserved position, a proline involved in a critical turn. Analysis of conserved P2 domains in a subset of Firmicutes predicted three of the nine

contact sites identified in the co-crystal, but conserved hydrophobic core residues were overpredicted as seen in the P2-III analysis. Conversely, the great evolutionary distance of the organisms in the family would seemingly make it possible to analyze potential P2-I interaction residues in their cognate CheY proteins. However, of the four subfamily specific residues conserved at 97% or more (with a background conservation level of 58% or less) in the CheY sequences predicted to interact with P2, none were located at the P2 interface. It turns out that the majority of the members of this CheY subfamily are also predicted to interact with CheC and these conserved residues more likely reflect that interaction. The low level of sequence conservation of P2 coupled with its lack of catalytic function suggests that residues maintaining the CheY-P2 interaction are able to diverge more easily than other interactions. The flexible binding modes seen in the CheY-P2-III co-crystal [20] further support this prediction.

5.2.2 CheC-CheY Contact Site Analysis

All CheC proteins are predicted to interact with CheY (Figure 3.12), while only a subset of CheY are predicted to interact with CheC. CheC conservation analysis showed only four positions conserved by 97% or more, and none of these positions included the experimentally identified active sites that were expected to be conserved given the enzymatic function of the protein. Phylogenetic analysis shows that some catalytic positions have deteriorated in CheC sequences from *Haloarcula marismortui*, *Halobacterium* sp. NRC-1, *Methanococcus maripaludis*, and *Natronomonas pharaonis*, all of which contain multiple CheC proteins that could be the cause of such divergence. When the CheC sequences from these organisms were removed, we were able to identify 13 residues conserved at 94% or higher. Three of the 13 residues are predicted to be buried, resulting in a final set of 10 residues important for CheC function and CheY interaction. All but one residue maps to the same face of CheC that includes both catalytic sites (Figure 5.7), which supports that they are involved in CheY interaction.

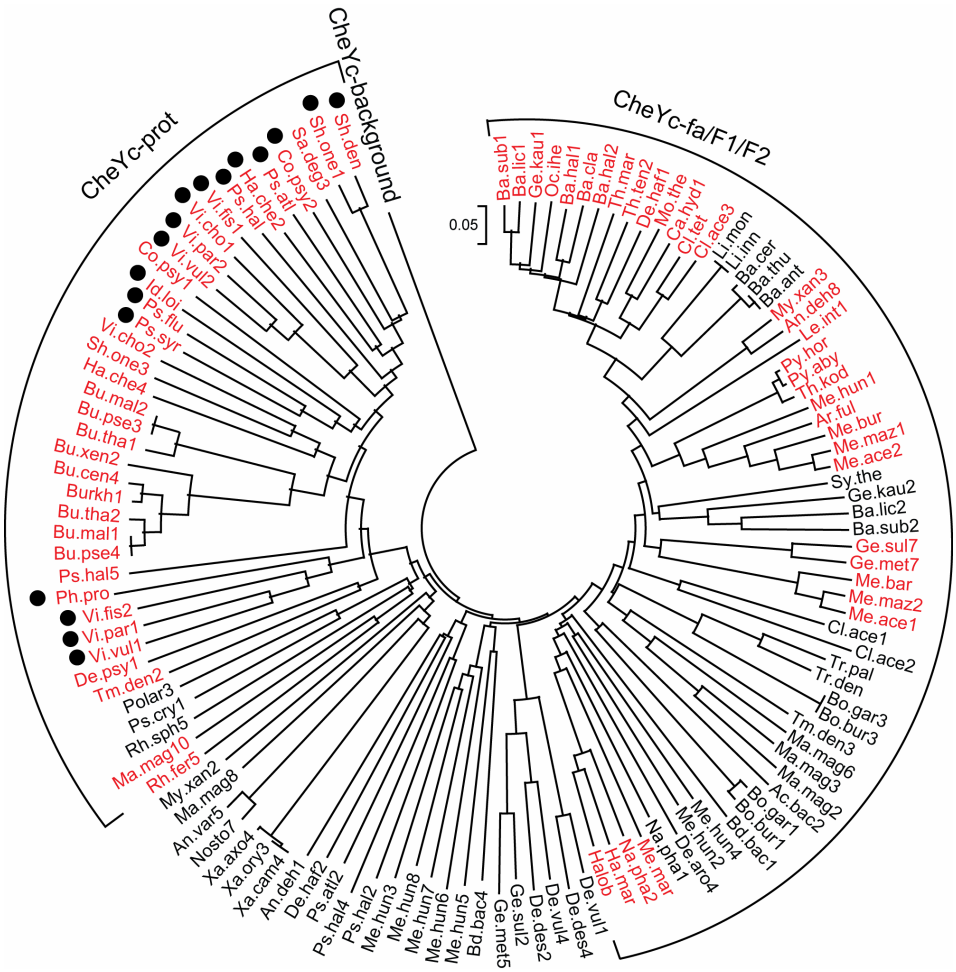


Figure 5.6 Like CheYz, the CheY tree shows a conserved family associated with the CheY-CheC interaction in Proteobacteria (CheYc-prot). The CheY-CheC interaction in Firmicutes and Archaea (CheYc-fa) is older and overlaps with the F1 and F2 CheY subfamilies. Black circles indicate CheY-CheC fusion sequences whose identifiers correspond to Table A.7. The remaining identifiers correspond to Table A.11. The red sequences in the CheY-prot set were included in consensus analysis since all are encoded next to or fused to CheY. The red sequences in the CheYc-fa were used in sequence analysis based on gene neighborhood and mirror tree analysis.

Gene neighborhood analysis shows that most CheC proteins are encoded adjacent to CheY or fused to CheY (Figure 3.12), but CheY phylogenetic analysis does not show the same clear grouping of a specific CheY family predicted to interact with CheC (CheYc) as seen in CheYz analysis. Instead we see two CheYc subfamilies; members of each subfamily that could be confirmed to interact with CheC based on gene neighborhood and mirror tree analyses were selected as the CheYc subfamily. Subfamily

analysis revealed only three CheYc specific residues. Given the divergence of the CheYc members seen in the CheY tree, we separated the CheYc set into two groups in accordance with the CheY phylogenetic analysis. One set is exclusively in Proteobacteria (CheYc-prot) and the other is primarily in Firmicutes and Archaea (CheYc-fa). We identified six CheYc-prot specific and six CheYc-fa specific residues after solvent accessibility analysis, which are conserved at 100% in their respective groups. Four positions are conserved in both sets, resulting in two positions uniquely conserved in each subset. One position that is conserved among both sets has a different physicochemical property in each set. All eight predicted contact site positions localize to the surface (Figure 5.7), seven of which are part of the same face associated with CheY phosphorylation that is most likely the site of CheC mediated dephosphorylation.

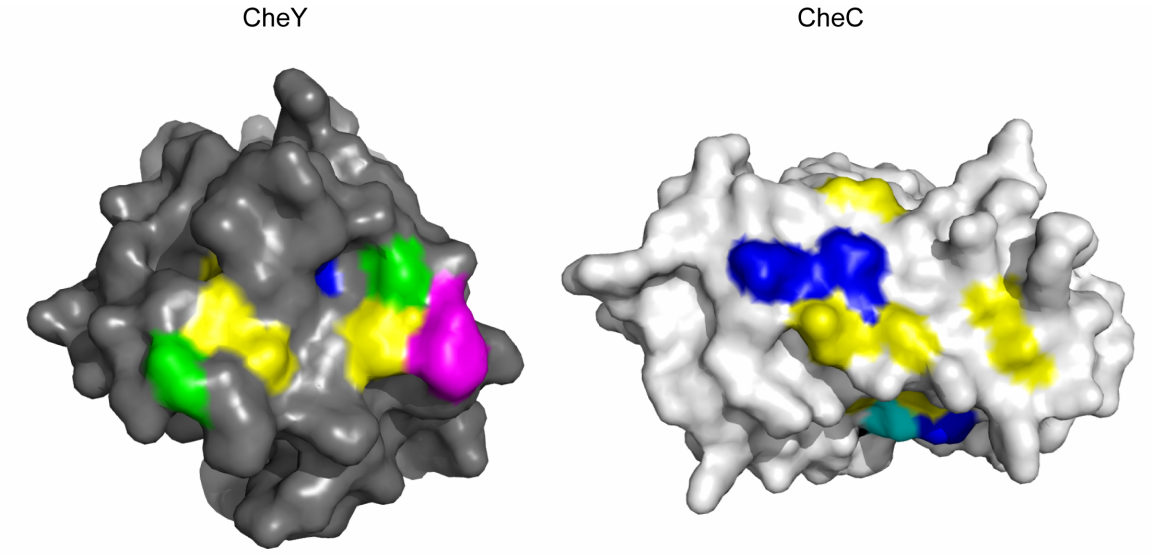


Figure 5.7 The predicted interaction faces of CheY and CheC. On CheY, positions specific to both CheYc data sets, the CheYc-fa data set only, and the CheYc-prot data set only are shown in yellow, pink, and green respectively as visualized on the CheY structure from *T. maritima*. The conserved site of phosphorylation (D54) is shown in blue for reference. The conserved CheC residues are visualized on the CheC structure from *T. maritima* in yellow and the conserved active site glutamate and asparagine residues are shown in dark blue. The light blue residue shows one active site asparagines that was not found to be conserved by 94% or more in our data set.

5.2.3 CheC-CheD Contact Site Analysis

Initial studies using the CheCd and CheDc subfamilies did not identify conserved subfamily specific residues in either group. These proteins are present in members of Firmicutes, Spirochete, Proteobacteria, and Archaea, unlike the relatively young CheYz subfamily that is only in Proteobacteria. In order to identify CheCd and CheDc specific residues we limited our CheCd and CheDc sets to a conserved subfamily of CheCd that contains Firmicutes and two laterally transferred members in Proteobacteria (CheCd-firm). The group also contains a member from the Spirochete, *Leptospira interrogans*, but it was excluded due to an unusually divergent cognate CheD in comparison to the remaining family members. The CheDc set (CheDc-firm) was identified as the CheDs encoded adjacent to the members of our CheCd group since the CheD tree shows poor taxonomic grouping in comparison to CheC (Figures 3.10 and 3.12). 52% and 51% background thresholds were used to identify CheCd-firm and CheDc-firm specific residues, respectively. CheC members not predicted to interact with CheD and vice versa were used as the background sets for CheCd-firm and CheDc-firm, respectively. Residues predicted to have 0% solvent accessibility in the CheC and CheD sequences from *T. maritima* were excluded from the final prediction set.

We identified six CheDc specific residues in the reduced CheDc-firm data set, four of which have sidechains within 3.6Å of any CheC atom and a glycine (G23 in *T. maritima*) that has been previously shown to be important to the interaction (Figure 5.8) [29]. The remaining residue identified by sequence analysis is located near the interface, though not within contact range, which suggests it could play an as yet unidentified role in the interaction. The six residues exclusively identified by distance are predicted to play only minor roles in maintaining the interaction. Although 13 CheC amino acid sidechains are within 3.6Å of any CheD atom, only four of these residues were specifically conserved in the CheCd-firm subfamily. Three other CheCd-firm specific

residues localized to the opposite side of CheC, away from the CheD interface, and the remaining two CheCd-firm specific residues are buried, but were not predicted as such in the solvent accessibility analysis. The three residues located on the opposite face localize to the same face associated with CheY interaction (Figure 5.7), which suggests they are specific CheY binding sites of the CheCd-firm subfamily. As defined in Table 5.1, the CheDc predictions showed a 0.833 specificity and 0.333 sensitivity, and the CheCd predictions showed a 0.444 specificity and a 0.286 sensitivity.

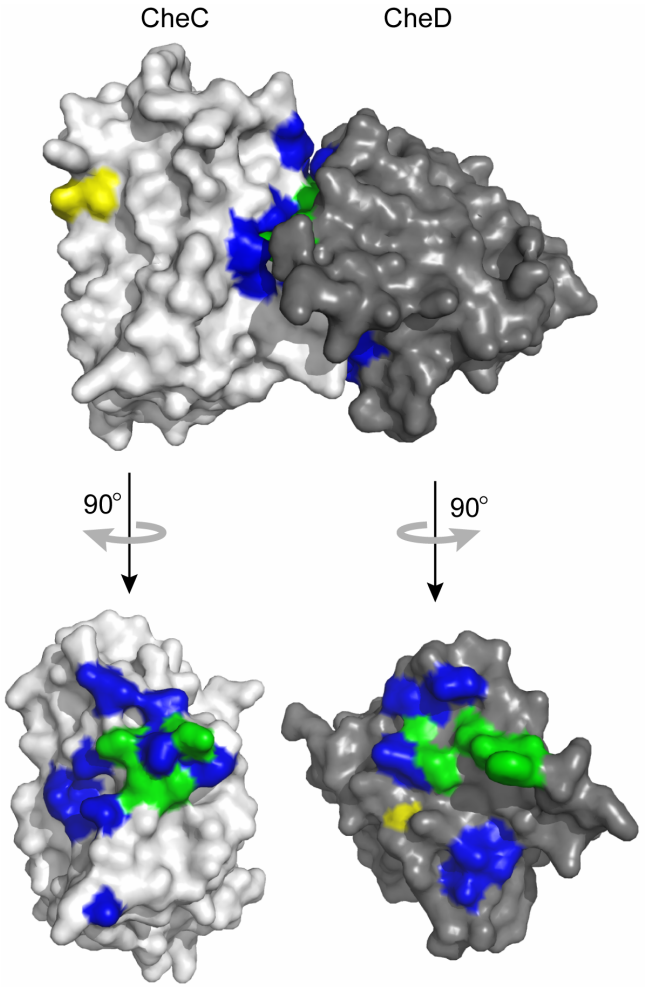


Figure 5.8 Contact site predictions for the CheC-CheD interaction. Residues involved in the interaction were identified as those with sidechains within 3.6Å of the partner protein. Contact sites exclusively predicted by structure analysis are shown in blue. Contact sites exclusively predicted by sequence analysis are shown in yellow. Contacts sites identified in both analyses are shown in green.

5.4 Conclusions

The work presented here is not a universal contact site prediction method, but instead, addresses core-accessory protein interaction scenarios. Our method has the advantage of exclusively using sequence based information since there is far more sequence data currently available than structural data. If individual structural information is available for one or both protein partners, it can only improve the performance of our method by allowing clear discrimination between surface and buried residues and conserved surface patches. The ability of our method to predict contact sites on CheZ, albeit at a reduced sensitivity, gives our method an advantage over some learning methods which excluded coiled-coils and homodimers from their analyzed data sets [240,246]. Although our method was used only to identify contact sites in a few proteins rather than hundreds, we find that a comparison to the learning methods is warranted since, unlike those methods, our method has the ability to identify residues that are still functionally involved in an interaction even if they are not true contact sites based on distance measures. Furthermore those methods only seek to identify residues that could be contact sites without discriminating between which protein or small molecule they contact, and this is especially problematic given that many proteins have more than one interaction partner. Learning methods do not require high quality multiple alignments, but the prevalence of multiple alignments in experimental publications suggests that they are not an unreasonable requirement of our method. When the CheYz, CheCd, and CheDc predictions were summed and validated only by structural data, the subfamily subtraction method shows a specificity of 0.6 and a sensitivity of 0.4, both of which are higher than the results of a high-throughput sequence-based learning method that utilizes support vector machines (SVMs) [247]. Our method also shows a higher specificity than another high-throughput sequence-based SVM method [246].

Correlated mutation analysis requires no structural data and has been used in both intraprotein [83,84,255] and interprotein [85,86] interactions. As with our method, it also requires high quality multiple alignments, but beyond that it also excludes analysis of paralog interactions, a situation that is common to many biological networks, because correlated mutation analysis requires a one-to-one relationship between interacting proteins [86]. Although that method has the advantage of predicting the specific residues that interact at the contact site, the high specificity of our residue conservation analysis allows for efficient analysis of interaction sites at the experimental level. Correlated mutation analysis of interprotein interactions is primarily applied to core-core scenarios, but it may be able to pick up less conserved core-accessory interaction sites that are not identified by our method. The CheY-CheZ interaction is relatively young for bacteria given that it is only present in Proteobacteria. For older interactions, we find that the method often maintains its high specificity as seen in the CheDc and CheCd analyses, with only minor data set alterations. In more mutable interactions such as that of CheY with P2, the method has greatly reduced specificity and sensitivity, but with greater restrictions on the identification of aliphatic residues it has the potential to be used for the identification of some interaction sites even in these cases.

In addition to the previously mentioned advantages, our method will be attractive to experimentalists studying the mechanisms of protein interactions for its simplicity and flexibility. Manual analysis allows it to be successful in predicting contact sites in vastly different types of protein-protein interactions. Our contact site prediction methods combine easy-to-use and available tools and provide more objective information than visual multiple alignment analysis. The thresholds can be varied on a case-by-case basis to best meet the needs of the users. Lowering the background percentage threshold for identifying subfamily specific residues within the core protein is likely to increase the specificity of the method, as is increasing the threshold for residues to be considered highly conserved in the subfamily set. The decreased background conservation level of

the highly conserved subfamily residues correlates to the decreased pressure to maintain their specific roles in structure and catalysis. This is a universal trend that supports its validity, potential for continued use, and possibility for automation. Although our method is not designed for quick analysis of many interactions, its ease of use and high specificity makes it convenient for case-by-case analysis in experimental environments.

APPENDIX

Table A.1 16S rRNA sources and location. ID1 corresponds to all protein identifiers associated with a particular organism in Tables A.2-16. ID2 corresponds to identifiers in the 16S phylogenetic tree (Figure 3.3). GI is the unique NCBI GenBank identifier for the genome sequence of the organism. The range is the location in the genome of the region used in our 16S multiple alignment. The strand corresponds to whether or not the region used is encoded on the positive (+) or negative (-) strand of the genome. A 16S minimum evolution tree was built in MEGA using complete deletion and the Tamura 3-parameter distance matrix.

ID1	ID2	Organism	GI	Range	Strand
Ac.bac	A.bacterium	Acidobacteria bacterium Ellin345	94967031	5260070-5261358	-
Acine	Acinetobacter sp. ADP1	Acinetobacter sp. ADP1	50083297	18630-19943	+
Ae.per	A.pernix	Aeropyrum pernix K1	14600379	1218712-1220677	-
Ag.tum	A.tumefaciens	Agrobacterium tumefaciens str. C58	15887359	56732-58021	+
An.deh	A.dehalogenans	Anaeromyxobacter dehalogenans 2CP-C	86156430	1361547-1362864	+
An.pha	A.phagocytophilum	Anaplasma phagocytophilum HZ	88606690	1057416-1058704	-
An.var	A.variabilis	Anabaena variabilis ATCC 29413	75906225	1003110-1004402	+
Aq.aeo	A.aeolicus	Aquifex aeolicus VF5	15282445	571203-572534	-
Ar.ful	A.fulgidus	Archaeoglobus fulgidus DSM 4304	11497621	1788988-1790257	-
As.yel	A.yellows	Aster yellows witches'-broom phytoplasma AYWB	85057280	271961-273262	+
Azoar	Azoarcus sp. EbN1	Azoarcus sp. EbN1	56475432	548902-550216	-
Ba.ant	B.anthracis	Bacillus anthracis str. Sterne	49183039	9543-10857	+
Ba.cer	B.cereus	Bacillus cereus ATCC 14579	30018278	9422-10736	+
Ba.cic	B.cicadellinicola	Baumannia cicadellinicola str. Hc (Homalodisca coagulata)	94676460	170988-172301	+
Ba.cla	B.clausii	Bacillus clausii KSM-K16	56961782	11909-13222	+
Ba.fra	B.fragilis	Bacteroides fragilis YCH46	53711291	3947855-3949150	-
Ba.hal	B.halodurans	Bacillus halodurans C-125	57596592	23051-24365	+
Ba.hen	B.henselae	Bartonella henselae str. Houston-1	49474831	1413133-1414422	-
Ba.lic	B.licheniformis	Bacillus licheniformis ATCC 14580	52783855	9939-11252	+
Ba.qui	B.quintana	Bartonella quintana str. Toulouse	49473688	1177325-1178614	-

Table A.1 (continued)

ID1	ID2	Organism	GI	Range	Strand
Ba.sub	B.subtilis	Bacillus subtilis subsp. subtilis str. 168	50812173	10042-11356	+
Ba.the	B.thetaiotaomicron	Bacteroides thetaiotaomicron VPI-5482	29345410	1630135-1631428	-
Ba.thu	B.thuringiensis	Bacillus thuringiensis serovar konkukian str. 97-27	49476684	9544-10858	+
Bd.bac	B.bacteriovorans	Bdellovibrio bacteriovorus HD100	42521650	819796-821086	+
Bi.lon	B.longum	Bifidobacterium longum NCC2705	58036264	159469-160767	+
Bo.bro	B.bronchiseptica	Bordetella bronchiseptica RB50	33598993	1973143-1974450	+
Bo.bur	B.burgdorferi	Borrelia burgdorferi B31	15594346	444584-445895	-
Bo.gar	B.garinii	Borrelia garinii PBi	51598263	446776-448087	-
Bo.par	B.parapertussis	Bordetella parapertussis 12822	33594723	1840771-1842078	+
Bo.per	B.pertussis	Bordetella pertussis Tohama I	33591275	2149264-2150571	+
Br.abo	B.abortus	Brucella abortus biovar 1 str. 9-941	62288991	1606419-1607708	-
Br.jap	B.japonicum	Bradyrhizobium japonicum USDA 110	27375111	1528423-1529712	+
Br.mel	B.melitensis	Brucella melitensis 16M	17986284	198840-200129	+
Br.sui	B.suis	Brucella suis 1330	56968325	1588604-1589893	-
Bu.aph	B.aphidicola	Buchnera aphidicola str. APS (Acyrthosiphon pisum)	15616630	274265-275581	+
Bu.cen	B.cenocepacia	Burkholderia cenocepacia AU 1054	107021562	57600-58907	+
Bu.mal	B.mallei	Burkholderia mallei ATCC 23344	53723370	1883677-1884984	-
Bu.pse	B.pseudomallei	Burkholderia pseudomallei K96243	53717639	1431989-1433296	+
Bu.tha	B.thailandensis	Burkholderia thailandensis E264	83716035	2494255-2495563	-
Bu.xen	B.xenovorans	Burkholderia xenovorans LB400	91777110	539724-541031	+
Burkh	Burkholderia sp. 383	Burkholderia sp. 383	78059643	139686-140993	+
Ca.Blo	C.Blochmannia	Candidatus Blochmannia floridanus	33519483	616530-617848	-
Ca.cre	C.crescentus	Caulobacter crescentus CB15	16124256	2840157-2841443	-
Ca.hyd	C.hydrogenoformans	Carboxydothermus hydrogenoformans Z-2901	78042616	1396198-1397519	-
Ca.jej	C.jejuni	Campylobacter jejuni subsp. jejuni NCTC 11168	15791399	39475-40760	+
Ca.Pel	C.Pelagibacter	Candidatus Pelagibacter ubique HTCC1062	71082709	511553-512829	+

Table A.1 (continued)

ID1	ID2	Organism	GI	Range	Strand
Ca.Pro	C.Protochlamydia	Candidatus Protochlamydia amoebophila UWE25	46445634	1230783-1232094	-
Ch.abo	C.abortus	Chlamydophila abortus S26/3	62184647	995245-996559	-
Ch.cav	C.caviae	Chlamydophila caviae GPIC	29839769	1023981-1025295	-
Ch.chl	C.chlorochromatii	Chlorobium chlorochromatii CaD3	78187984	102814-104092	+
Ch.fel	C.felis	Chlamydophila felis Fe/C-56	89897807	147934-149248	+
Ch.mur	C.muridarum	Chlamydia muridarum Nigg	29337300	134071-135385	+
Ch.pne	C.pneumoniae	Chlamydophila pneumoniae CWL029	15617929	1000794-1002110	+
Ch.sal	C.salexigens	Chromohalobacter salexigens DSM 3043	92112136	455137-456447	+
Ch.tep	C.tepidum	Chlorobium tepidum TLS	21672841	139707-140985	+
Ch.tra	C.trachomatis	Chlamydia trachomatis D/UW-3/CX	15604717	854359-855673	+
Ch.vio	C.violaceum	Chromobacterium violaceum ATCC 12472	34495455	421444-422756	+
Cl.ace	C.acetobutylicum	Clostridium acetobutylicum ATCC 824	15893298	9927-11214	+
Cl.per	C.perfringens	Clostridium perfringens str. 13	18308982	10392-11678	+
Cl.tet	C.tetani	Clostridium tetani E88	28209834	8720-10004	-
Co.bur	C.burnetii	Coxiella burnetii RSA 493	77358712	165801-167113	+
Co.dip	C.diphtheriae	Corynebacterium diphtheriae NCTC 13129	38232642	743575-744870	+
Co.eff	C.efficiens	Corynebacterium efficiens YS-314	25026556	2989974-2991270	-
Co.glu	C.glutamicum	Corynebacterium glutamicum ATCC 13032	58036263	76864-78161	+
Co.jei	C.jeikeium	Corynebacterium jeikeium K411	68535062	108906-110206	+
Co.psy	C.psychrerythraea	Colwellia psychrerythraea 34H	71277742	35134-36449	+
De.aro	D.aromatica	Dechloromonas aromatica RCB	71905642	77064-78378	+
De.des	D.desulfuricans	Desulfovibrio desulfuricans G20	78355047	70074-71390	+
De.eth	D.ethenogenes	Dehalococcoides ethenogenes 195	57233530	928407-929685	-
De.geo	D.geothermalis	Deinococcus geothermalis DSM 11300	94984109	400522-401816	+
De.haf	D.hafniense	Desulfitobacterium hafniense Y51	89892746	237336-238655	+
De.psy	D.psychrophila	Desulfotalea psychrophila LSv54	51243852	806440-807757	+
De.rad	D.radiodurans	Deinococcus radiodurans R1	15805042	85043-86334	+

Table A.1 (continued)

ID1	ID2	Organism	GI	Range	Strand
De.vul	D.vulgaris	Desulfovibrio vulgaris subsp. vulgaris str. Hildenborough	46562128	106124-107439	+
Dehal	Dehalococcoides sp. CBDB1	Dehalococcoides sp. CBDB1	73747956	803544-804822	-
Eh.can	E.canis	Ehrlichia canis str. Jake	73666633	286150-287437	+
Eh.cha	E.chaffeensis	Ehrlichia chaffeensis str. Arkansas	88657561	942222-943508	-
Eh.rum	E.ruminantium	Ehrlichia ruminantium str. Welgevonden	57238731	327153-328439	+
En.fae	E.faecalis	Enterococcus faecalis V583	29374661	248703-250016	+
Er.car	E.carotovora	Erwinia carotovora subsp. atroseptica SCRI1043	50118965	3757996-3759309	-
Er.lit	E.litoralis	Erythrobacter litoralis HTCC2594	85372828	63725-65013	+
Es.col	E.coli	Escherichia coli O157:H7 EDL933	16445223	227329-228641	+
Fr.tul	F.tularensis	Francisella tularensis subsp. tularensis Schu 4	56707187	1310943-1312253	-
Frank	Frankia sp. CcI3	Frankia sp. CcI3	86738724	3751634-3752928	-
Fu.nuc	F.nucleatum	Fusobacterium nucleatum subsp. nucleatum ATCC 25586	19703352	451479-452770	-
Ge.kau	G.kaustophilus	Geobacillus kaustophilus HTA426	56418535	10650-11968	+
Ge.met	G.metallireducens	Geobacter metallireducens GS-15	78221228	1310272-1311589	+
Ge.sul	G.sulfurreducens	Geobacter sulfurreducens PCA	39995111	684926-686243	+
Gl.oxy	G.oxydans	Gluconobacter oxydans 621H	58038491	241940-243224	-
Gl.vio	G.violaceus	Gloeobacter violaceus PCC 7421	37519569	1571234-1572522	-
Ha.che	H.chejuensis	Hahella chejuensis KCTC 2396	83642913	1588565-1589878	+
Ha.duc	H.ducreyi	Haemophilus ducreyi 35000HP	33151282	10868-12179	+
Ha.inf	H.influenzae	Haemophilus influenzae Rd KW20	16271976	127179-128489	-
Ha.mar	H.marismortui	Haloarcula marismortui ATCC 43049	55376942	1106974-1108239	-
Halob	Halobacterium sp. NRC-1	Halobacterium sp. NRC-1	15789340	1875710-1876976	+
He.aci	H.acinonychis	Helicobacter acinonychis str. Sheeba	109946640	152013-153301	+
He.hep	H.hepaticus	Helicobacter hepaticus ATCC 51449	32265499	956843-958125	-
He.pyl	H.pylori	Helicobacter pylori J99	15611071	1188030-1189318	-

Table A.1 (continued)

ID1	ID2	Organism	GI	Range	Strand
Id.loi	I.loihiensis	Idiomarina loihiensis L2TR	56459112	1266667-1267981	-
Janna	Jannaschia sp. CCS1	Jannaschia sp. CCS1	89052491	4083220-4084488	-
La.aci	L.acidophilus	Lactobacillus acidophilus NCFM	58336354	59502-60814	+
La.del	L.delbrueckii	Lactobacillus delbrueckii subsp. bulgaricus ATCC 11842	104773257	1787065-1788379	-
La.int	L.intracellularis	Lawsonia intracellularis PHE/MN1-00	94986445	675001-676316	-
La.joh	L.johnsonii	Lactobacillus johnsonii NCC 533	42518084	558836-560150	+
La.lac	L.lactis	Lactococcus lactis subsp. lactis Il1403	15671982	537794-539105	+
La.pla	L.plantarum	Lactobacillus plantarum WCFS1	28376974	503115-504428	+
La.sak	L.sakei	Lactobacillus sakei subsp. sakei 23K	81427616	306428-307742	+
La.sal	L.salivarius	Lactobacillus salivarius subsp. salivarius UCC118	90960990	74759-76071	+
Le.int	L.interrogans	Leptospira interrogans serovar Lai str. 56601	24212700	2417575-2418869	-
Le.pne	L.pneumophila	Legionella pneumophila subsp. pneumophila str. Philadelphia 1	52840256	360087-361401	+
Le.xyl	L.xyli	Leifsonia xyli subsp. xyli str. CTCB07	50953925	134012-135308	+
Li.inn	L.innocua	Listeria innocua Clip11262	16799079	260764-262076	+
Li.mon	L.monocytogenes	Listeria monocytogenes EGD-e	16802048	237703-239015	+
Ma.mag	M.magneticum	Magnetospirillum magneticum AMB-1	83309099	965097-966383	+
Ma.suc	M.succiniciproducens	Mannheimia succiniciproducens MBEL55E	52424055	149760-151072	+
Me.ace	M.acetivorans	Methanosarcina acetivorans C2A	20088899	1073249-1074517	+
Me.bar	M.barkeri	Methanosarcina barkeri str. fusaro	73667559	1141482-1142750	-
Me.bur	M.burtonii	Methanococcoides burtonii DSM 6242	91772082	1022126-1023393	-
Me.cap	M.capsulatus	Methylococcus capsulatus str. Bath	77128441	788364-789677	+
Me.fla	M.flagellatus	Methylobacillus flagellatus KT	91774356	64334-65649	+
Me.flo	M.florum	Mesoplasma florum L1	50364815	192071-193372	+
Me.hun	M.hungatei	Methanospirillum hungatei JF-1	88601322	40017-41278	+
Me.jan	M.jannaschii	Methanocaldococcus jannaschii DSM 2661	15668172	157985-159247	-

Table A.1 (continued)

ID1	ID2	Organism	GI	Range	Strand
Me.kan	M.kandleri	Methanopyrus kandleri AV19	20093440	516777-518055	-
Me.lot	M.loti	Mesorhizobium loti MAFF303099	57165207	2750004-2751293	-
Me.mar	M.maripaludis	Methanococcus maripaludis S2	45357563	3354-4618	-
Me.maz	M.mazei	Methanosarcina mazei Go1	21226102	235307-236575	-
Me.sta	M.stadtmanae	Methanosphaera stadtmanae DSM 3091	84488831	408871-410141	+
Me.the	M.thermautotrophicus	Methanothermobacter thermautotrophicus str. Delta H	15678031	1602417-1603687	+
Mo.the	M.thermoacetica	Moorella thermoacetica ATCC 39073	83588874	148235-149552	+
My.avi	M.avium	Mycobacterium avium subsp. paratuberculosis K-10	41406098	2751286-2752595	-
My.bov	M.bovis	Mycobacterium bovis AF2122/97	31791177	1470107-1471413	+
My.cap	M.capricolum	Mycoplasma capricolum subsp. capricolum ATCC 27343	83319253	464650-465949	+
My.gal	M.gallisepticum	Mycoplasma gallisepticum R	31544204	320265-321558	+
My.gen	M.genitalium	Mycoplasma genitalium G37	108885074	170233-171526	+
My.hyo	M.hyopneumoniae	Mycoplasma hyopneumoniae 232	54019969	876249-877546	-
My.lep	M.leprae	Mycobacterium leprae TN	15826865	1341380-1342687	+
My.mob	M.mobile	Mycoplasma mobile 163K	47458835	646165-647461	-
My.myc	M.mycoides	Mycoplasma mycoides subsp. mycoides SC str. PG1	42560560	678212-679510	-
My.pen	M.penetrans	Mycoplasma penetrans HF-2	26553452	1236494-1237785	-
My.pne	M.pneumoniae	Mycoplasma pneumoniae M129	13507739	118535-119826	+
My.pul	M.pulmonis	Mycoplasma pulmonis UAB CTIP	15828471	813588-814885	-
My.syn	M.synoviae	Mycoplasma synoviae 53	71894025	665113-666395	-
My.tub	M.tuberculosis	Mycobacterium tuberculosis CDC1551	50953765	1471612-1472918	+
My.xan	M.xanthus	Myxococcus xanthus DK 1622	108756767	376677-377975	+
Mycob	Mycobacterium sp. MCS	Mycobacterium sp. MCS	108796981	3116985-3118283	-
Na.equ	N.equitans	Nanoarchaeum equitans Kin4-M	38349555	432557-433824	+
Na.pha	N.pharaonis	Natronomonas pharaonis DSM 2160	76800655	214464-215728	-

Table A.1 (continued)

ID1	ID2	Organism	GI	Range	Strand
Ne.gon	N.gonorrhoeae	Neisseria gonorrhoeae FA 1090	59800473	1119760-1121071	-
Ne.men	N.meningitidis	Neisseria meningitidis Z2491	15793034	198345-199656	-
Ne.sen	N.sennetsu	Neorickettsia sennetsu str. Miyayama	88607955	36476-37762	+
Ni.eur	N.europaea	Nitrosomonas europaea ATCC 19718	30248031	69358-70673	+
Ni.ham	N.hamburgensis	Nitrobacter hamburgensis X14	92115633	723825-725114	+
Ni.mul	N.multiformis	Nitrosospira multiformis ATCC 25196	82701135	562842-564154	+
Ni.oce	N.oceani	Nitrosococcus oceani ATCC 19707	77163561	999604-1000917	+
Ni.win	N.winogradskyi	Nitrobacter winogradskyi Nb-255	75674199	638499-639788	+
No.aro	N.aromaticivorans	Novosphingobium aromaticivorans DSM 12444	87198026	2851033-2852320	-
No.far	N.farcinica	Nocardia farcinica IFM 10152	54021964	1179773-1181070	+
Nosto	Nostoc sp. PCC 7120	Nostoc sp. PCC 7120	17227497	2375927-2377219	+
Oc.ihe	O.iheyensis	Oceanobacillus iheyensis HTE831	23097455	91741-93053	+
On.yel	O.yellows	Onion yellows phytoplasma OY-M	39938486	279622-280923	+
Pa.mul	P.multocida	Pasteurella multocida subsp. multocida str. Pm70	15601865	1080343-1081654	-
Pe.car	P.carbinolicus	Pelobacter carbinolicus DSM 2380	90960985	2750587-2751903	-
Pe.lut	P.luteolum	Pelodictyon luteolum DSM 273	78185892	107914-109192	+
Ph.lum	P.luminescens	Photorhabdus luminescens subsp. laumondii TTO1	37524032	58666-59979	+
Ph.pro	P.profundum	Photobacterium profundum SS9	54301680	2011036-2012349	-
Pi.tor	P.torridus	Picrophilus torridus DSM 9790	48477072	470717-471982	-
Po.gin	P.gingivalis	Porphyromonas gingivalis W83	34539880	119787-121084	+
Polar	Polaromonas sp. JS666	Polaromonas sp. JS666	91785913	4452036-4453347	-
Pr.acn	P.acnes	Propionibacterium acnes KPA171202	50841496	606380-607677	+
Pr.mar	P.marinus	Prochlorococcus marinus subsp. marinus str. CCMP1375	33239452	353502-354792	+
Ps.aer	P.aeruginosa	Pseudomonas aeruginosa PAO1	110645304	722316-723628	+
Ps.arc	P.arcticus	Psychrobacter arcticus 273-4	71064581	660770-662083	+

Table A.1 (continued)

ID1	ID2	Organism	GI	Range	Strand
Ps.atl	P.atlantica	Pseudoalteromonas atlantica T6c	109896332	36302-37613	+
Ps.cry	P.cryohalolentis	Psychrobacter cryohalolentis K5	93004831	652225-653538	+
Ps.ent	P.entomophila	Pseudomonas entomophila L48	104779316	115712-117025	+
Ps.flu	P.fluorescens	Pseudomonas fluorescens Pf-5	70728250	123033-124346	+
Ps.hal	P.haloplanktis	Pseudoalteromonas haloplanktis TAC125	77358982	35353-36665	+
Ps.put	P.putida	Pseudomonas putida KT2440	26986745	171592-172905	+
Ps.syr	P.syringae	Pseudomonas syringae pv. tomato str. DC3000	28867243	666949-668262	+
Py.aby	P.abyssi	Pyrococcus abyssi GE5	14518450	205274-206546	+
Py.aer	P.aerophilum	Pyrobaculum aerophilum str. IM2	18311643	1089869-1091851	+
Py.fur	P.furiosus	Pyrococcus furiosus DSM 3638	18976372	137153-138425	+
Py.hor	P.horikoshii	Pyrococcus horikoshii OT3	14589963	191198-192470	+
Ra.eut	R.eutropha	Ralstonia eutropha JMP134	73537298	180552-181860	+
Ra.met	R.metallidurans	Ralstonia metallidurans CH34	94308945	3431196-3432504	-
Ra.sol	R.solanacearum	Ralstonia solanacearum GMI1000	17544719	1532922-1534230	+
Rh.bal	R.baltica	Rhodopirellula baltica SH 1	32470666	5076964-5078281	-
Rh.etl	R.etli	Rhizobium etli CFN 42	86355669	62782-64067	+
Rh.fer	R.ferrireducens	Rhodoferax ferrireducens T118	89898822	4112644-4113955	-
Rh.pal	R.palustris	Rhodopseudomonas palustris CGA009	39933080	4996221-4997510	-
Rh.rub	R.rubrum	Rhodospirillum rubrum ATCC 11170	83591340	192715-194001	+
Rh.sph	R.sphaeroides	Rhodobacter sphaeroides 2.4.1	77461965	195-1462	+
Ri.bel	R.bellii	Rickettsia bellii RML369-C	91204815	538013-539300	+
Ri.con	R.conorii	Rickettsia conorii str. Malish 7	15891923	884604-885891	-
Ri.fel	R.felis	Rickettsia felis URRWXCal2	67458392	456600-457887	+
Ri.pro	R.prowazekii	Rickettsia prowazekii str. Madrid E	15603881	772266-773552	-
Ri.typ	R.typhi	Rickettsia typhi str. Wilmington	51473215	779672-780959	-
Ru.xyl	R.xylanophilus	Rubrobacter xylanophilus DSM 9941	108802856	1337570-1338893	+
Sa.deg	S.degradans	Saccharophagus degradans 2-40	90019649	1410054-1411367	+

Table A.1 (continued)

ID1	ID2	Organism	GI	Range	Strand
Sa.ent	S.enterica	Salmonella enterica subsp. enterica serovar Typhi str. CT18	16758993	287705-289017	+
Sa.rub	S.ruber	Salinibacter ruber DSM 13855	83814055	3328512-3329819	-
Sh.boy	S.boydii	Shigella boydii Sb227	82542618	209973-211285	+
Sh.den	S.denitrificans	Shewanella denitrificans OS217	91791369	42815-44128	+
Sh.dys	S.dysenteriae	Shigella dysenteriae Sd197	82775382	224970-226282	+
Sh.fle	S.flexneri	Shigella flexneri 2a str. 2457T	30061571	2720029-2721341	-
Sh.one	S.oneidensis	Shewanella oneidensis MR-1	24371600	46333-47646	+
Sh.son	S.sonnei	Shigella sonnei Ss046	74310614	240310-241622	+
Si.mel	S.meliloti	Sinorhizobium meliloti 1021	15963753	81959-83248	+
Si.pom	S.pomeroyi	Silicibacter pomeroyi DSS-3	56694928	262097-263362	+
Silic	Silicibacter sp. TM1040	Silicibacter sp. TM1040	99079841	145149-146413	+
So.glo	S.glossinidius	Sodalis glossinidius str. 'morsitans'	85057978	200470-201782	+
Sp.ala	S.alaskensis	Sphingopyxis alaskensis RB2256	103485498	2851571-2852862	-
St.aga	S.agalactiae	Streptococcus agalactiae 2603V/R	22536185	16636-17949	+
St.aur	S.aureus	Staphylococcus aureus subsp. aureus MW2	21281729	491768-493083	+
St.ave	S.avermitilis	Streptomyces avermitilis MA-4680	57833846	3077213-3078517	-
St.coe	S.coelicolor	Streptomyces coelicolor A3(2)	32141095	1472198-1473502	-
St.epi	S.epidermidis	Staphylococcus epidermidis ATCC 12228	27466918	1598011-1599325	-
St.hae	S.haemolyticus	Staphylococcus haemolyticus JCSC1435	70725001	880068-881382	+
St.mut	S.mutans	Streptococcus mutans UA159	24378532	17110-18423	+
St.pne	S.pneumoniae	Streptococcus pneumoniae TIGR4	15899949	15578-16890	+
St.pyo	S.pyogenes	Streptococcus pyogenes M1 GAS	15674250	17298-18611	+
St.sap	S.saprophyticus	Staphylococcus saprophyticus subsp. saprophyticus ATCC 15305	73661309	743949-745265	+
St.the	S.thermophilus	Streptococcus thermophilus LMG 18311	55820103	18050-19362	+
Su.aci	S.acidocaldarius	Sulfolobus acidocaldarius DSM 639	70605853	1107136-1108397	-
Su.sol	S.solfataricus	Sulfolobus solfataricus P2	15896971	871904-873168	+

Table A.1 (continued)

ID1	ID2	Organism	GI	Range	Strand
Su.tok	S.tokodaii	Sulfolobus tokodaii str. 7	24473558	1186505-1187768	-
Sy.aci	S.aciditrophicus	Syntrophus aciditrophicus SB	85857845	738248-739570	+
Sy.elo	S.elongatus	Synechococcus elongatus PCC 6301	56750010	1050804-1052096	-
Sy.the	S.thermophilum	Symbiobacterium thermophilum IAM 14863	51891138	21384-22707	+
Synco1	Synechococcus sp. WH 8102	Synechococcus sp. WH 8102	33864539	1874660-1875947	-
Synco2	Synechococcus sp. JA-3-3Ab	Synechococcus sp. JA-3-3Ab	86604733	1109997-1111269	-
Syncy	Synechocystis sp. PCC 6803	Synechocystis sp. PCC 6803	16329170	2452057-2453497	-
Tb.den	Tb.denitrificans	Thiobacillus denitrificans ATCC 25259	74316018	478292-479601	+
Th.aci	T.acidophilum	Thermoplasma acidophilum DSM 1728	16081186	1474300-1475565	-
Th.elo	T.elongatus	Thermosynechococcus elongatus BP-1	22297544	2335247-2336539	-
Th.fus	T.fusca	Thermobifida fusca YX	72160406	3594544-3595855	-
Th.kod	T.kodakarensis	Thermococcus kodakarensis KOD1	57639935	2023078-2024351	+
Th.mar	T.maritima	Thermotoga maritima MSB8	15642775	189198-190522	+
Th.ten	T.tengcongensis	Thermoanaerobacter tengcongensis MB4	20806542	1488417-1489803	-
Th.the	T.thermophilus	Thermus thermophilus HB27	46198308	1310198-1311492	-
Th.vol	T.volcanium	Thermoplasma volcanium GSS1	13540831	1518774-1520039	-
Tm.cru	T.crunogena	Thiomicrospira crunogena XCL-2	78484346	1635980-1637292	-
Tm.den	Tm.denitrificans	Thiomicrospira denitrificans ATCC 33889	78776201	453325-454615	+
Tr.den	T.denticola	Treponema denticola ATCC 35405	42516522	610411-611725	+
Tr.pal	T.pallidum	Treponema pallidum subsp. pallidum str. Nichols	15638995	230331-231645	+
Tr.whi	T.whipplei	Tropheryma whipplei TW08/27	28572175	680289-681583	-
Ur.par	U.parvum	Ureaplasma parvum serovar 3 str. ATCC 700970	13357558	145562-146852	+
Vi.cho	V.cholerae	Vibrio cholerae O1 biovar eltor str. N16961	15640032	54042-55355	+
Vi.fis	V.fischeri	Vibrio fischeri ES114	59710607	347484-348797	+
Vi.par	V.parahaemolyticus	Vibrio parahaemolyticus RIMD 2210633	28896774	33867-35180	+
Vi.vul	V.vulnificus	Vibrio vulnificus CMCP6	27363490	978404-979717	-

Table A.1 (continued)

ID1	ID2	Organism	GI	Range	Strand
Wi.glo	W.glossinidia	Wigglesworthia glossinidia endosymbiont of Glossina brevipalpis	32490749	136818-138129	+
Wo.end	W.endosymbiont	Wolbachia endosymbiont of Drosophila melanogaster	42519920	1168148-1169437	+
Wo.suc	W.succinogenes	Wolinella succinogenes DSM 1740	34556458	136396-137680	+
Xa.axo	X.axonopodis	Xanthomonas axonopodis pv. citri str. 306	21240774	4580058-4581373	-
Xa.cam	X.campestris	Xanthomonas campestris pv. campestris str. ATCC 33913	21229478	4561298-4562613	-
Xa.ory	X.oryzae	Xanthomonas oryzae pv. oryzae KACC10331	58579623	305957-307272	+
Xy.fas	X.fastidiosa	Xylella fastidiosa 9a5c, complete genome	57014152	66784-68099	+
Ye.pes	Y.pestis	Yersinia pestis CO92	16120353	12491-13804	+
Ye.pse	Y.pseudotuberculosis	Yersinia pseudotuberculosis IP 32953	51594359	3509149-3510462	-
Zy.mob	Z.mobilis	Zymomonas mobilis subsp. mobilis ZM4	56550896	1616555-1617846	-

Table A.2 CheA data. IDs (minus the numbers) correspond to the organisms from Table A.1. GI is the NCBI Genbank identifier. The gene neighborhood corresponds to the *cheA* (A), *cheB* (B), *cheC* (C), *cheD* (D), *cheR* (R), *cheV* (V), *cheW* (W), *cheY* (Y), *cheZ* (Z), and *mcp* (M) neighboring the region encoding the sequence. Any other gene in the neighborhood that is not one of the aforementioned is labeled with a dash (-). Numbers in the gene neighborhoods correspong to specific sequences in each table (i.e. Ac.bac1 corresponds to A1 and the B1 in its neighborhood corresponds to Ac.bac1 in Table A.3). There are no numbers for the *mcps* since they are too numerous to be manageable in such a format. Capitalized genes are encoded from left to right, and lower case genes are encoded from right to left. Class signifies the subfamilies identified in Figure 3.7. Range corresponds to the location of the core region of the protein (if present) used in the multiple alignment and phylogenetic analysis (the P3-P5 domains in this case). The location of the P2 domains in the sequences is listed if they are present. One asterisk (*) marks sequences from organisms without sequenced genomes that were included in the analysis due to available experimental information. Gene neighborhoods are taken from the experimental information. Two asterisks (**) mark sequences of split CheA proteins that were excluded from the P3-P5 phylogenetic analysis. Three asterisks (***) mark a divergent CheA that lacks the P3 and P4 domains, leading it to be excluded from the P3-P5 analysis. A maximum likelihood tree of the CheA core was built in PHYML with four substitution rate categories that have a fixed gamma distribution parameter of 2.00.

ID	GI	Gene Neighborhood	Class	Range	P2
Ac.bac1	94968552	A1W1MY1B1R1	F5	220-598	
Ac.bac2	94968798	MW2R2A2B2Y2X1	F2	323-707	163-241
Acine	50084005	Y1Y2W-A	Tfp	831-1283	
Ag.tum	15887865	M-Y1ARBY2D-----M	F7a	360-738	183-268
An.deh1	86157031	r2b2a2mw3w2Y3--Y2-D1A1W1R1B1	F8	163-537	
An.deh2	86157041	b1r1w1a1d1-y2--y3W2W3MA2B2R2	Alt	187-607	
An.deh3	86157619	R3W4MA3B3	Alt	138-564	
An.deh4	86157796	A4R4D2b4mw5--Y5A5MW6R5B5	F7a	291-667	
An.deh5	86157805	A4R4D2b4mw5--Y5A5MW6R5B5	F8	278-649	154-239
An.deh6	86159158	A6-W9W8-R6B6Y7	F10	291-673	175-248
An.deh7	86160731	MW11Y8A7C	F1	295-672	185-263
An.var1	75906287	Y1Y2W1MA1	Tfp	338-807	
An.var2	75906728	Y4Y3W2MA2	Tfp	1687-2162	
An.var3	75909983	Y6Y7W3MA3	Tfp	522-998	
Ar.ful	11498645	MW-YBACDR--M	F1a	276-649	145-224
Az.bra*	17864025	AWYBR	F5	449-830	
Azoar	56476385	Y1Y2WMA	Tfp	1173-1624	
Ba.ant	49184558	YAA---r2	F1b	87-468	1-28
Ba.ant-N**	49184557	YAA---r2	F1b		146-199
Ba.cer	30019775	YA---r2	F1b	291-672	147-227
Ba.cla	56964347	AW	F1a	271-645	157-237
Ba.hal	15615532	AWY2	F1a	305-680	175-255
Ba.lic	52080246	BAWCD	F1a	296-671	166-246
Ba.sub	16078706	BAWCD	F1a	295-669	166-245
Ba.thu	49477310	YA---r2	F1b	286-667	147-227

Table A.2 (continued)

ID	GI	Gene Neighborhood	Class	Range	P2
Bd.bac1	42522173	A1W1D1/B1		176-563	
Bd.bac2	42524822	w3-A2R2B2	F7a	264-659	
Bo.bro	33601525	Y1AWMRBY2Z	F7b	295-671	171-237
Bo.bur1	15594912	W2-A1B2-Y2	F8	333-709	180-265
Bo.bur2	15595014	A2R3/W3XY3	F2	414-855	299-378
Bo.gar1	51598818	W2-A1B2-Y2	F8	332-708	179-264
Bo.gar2	51598924	A2R3/W3XY3	F2	423-864	308-387
Bo.par	33596126	Y1AWMRBY2Z	F7b	293-669	171-237
Bo.per	33592173	AWMBYZ	F7b	293-669	171-237
Br.jap1	27375504	A1W1Y1R1	F5	269-648	
Br.jap2	27377303	A2W2Y2B1R2	F5	269-648	
Br.jap3	27377454	Y3A3W3MW4MR3B2	F8	249-619	95-180
Bu.cen1	107024394	Y2A1W1MR1DB1Y1Z	F7b	361-737	192-258
Bu.cen2	107026915	MW3R2-W2A2B3	Alt	193-617	
Bu.mal1	53717516	MW2R1W1A1B2	Alt	290-714	
Bu.mal2	53724322	Y4A2W4MR2DB1-Y3Z	F7b	372-748	209-275
Bu.pse1	53720916	Y2A1W2MR1DB1Y1Z	F7b	352-728	191-257
Bu.pse2	53722891	MW4R2W3A2B2	Alt	290-714	
Bu.tha1	83716091	Y3A1W2MR1D1B2B1	F7b	357-731	180-248
Bu.tha2	83716826	MW1R2W3A2B3	Alt	258-682	
Bu.tha3	83719847	Y4A3W4MR3D2B4Y5Z	F7b	354-730	189-255
Bu.xen	91785658	Y4AWMR4DB4Y3Z	F7b	376-752	201-267
Burkh1	78063151	MW2R1W1A1B2	Alt	193-617	
Burkh2	78064835	Y3A2W3MR2DB3Y4Z	F7b	354-730	191-257
Ca.cre1	16124688	M-M-Y1A1W1R1B1Y2D	F7a	362-740	192-277
Ca.cre2	16124848	mzy3-MA2W2Y4B2R2	F5	215-590	
Ca.hyd1	78042810	MW2A1B1	F1	301-685	164-244
Ca.hyd2	78044758	DR2CY1----M-MY3-R1MW1B2Y2A2	F9	518-895	171-257
Ca.jej	15791654	VAW	F3	235-618	
Ch.sal	92114146	AWMRBMYZ	F7b	313-689	179-245
Ch.vio1	34496469	Y1-A1MW1MD1B1	F8	334-704	150-236
Ch.vio2	34497965	MY2-A2MW2R2B3	F8	329-697	161-247
Ch.vio3	34498897	a4zy5v3v2--Y4A3W3--MR3B5D2	F7b	353-729	152-222
Ch.vio4	34498905	d2b5r3m--w3a3y4--V2V3Y5ZA4	F6	236-614	
Cl.ace1	15893414	Y1A1W1MR1Y2	F7a	262-641	140-221
Cl.ace2	15895488	W3DBR2A2CY3W2	F1a	323-692	181-261
Cl.tet	28211383	W2DBRACYW1	F1a	428-797	293-373
Co.psy	71277814	Y2ZAB---W2W1	F6	323-698	
De.aro1	71906365	MY1A1W1MMR1D1B1-V1V2Y2ZA2	F7b	396-772	157-233
De.aro2	71906377	MY1A1W1MMR1D1B1-V1V2Y2ZA2	F6	241-619	
De.aro3	71906781	Y3M-A3MW3D2R2D3B2	F8	347-714	164-250
De.aro4	71909506	Y7Y6W4MA4	Tfp	1261-1712	
De.des1	78356620	B1-R2-W3Y3A1	F4a	560-945	
De.des2	78357149	MY5A2R3B2	F8	332-701	165-250
De.haf	89895740	m--AW2B	F1a	330-711	176-255

Table A.2 (continued)

ID	GI	Gene Neighborhood	Class	Range	P2
De.psy	51246494	AW2B3R2---x1x2y2	F5	239-628	
De.rad	15808040	YWMMMA	Tfp	36-494	
De.vul1	46580005	Y2A1R2B2	F8	329-699	154-239
De.vul2	46580369	MW3A2	F8	314-683	167-251
De.vul3	46580477	B3-R3-W4Y3A3	F4a	695-1080	
Er.car	50120623	m-----AWMRBYZ	F7b	276-652	158-224
Er.lit	85375080	AWYBR-M	F5	151-529	
Es.col	15802300	AWMMRBYZ	F7b	264-640	160-226
Ge.kau	56419777	BAW1CD	F1a	289-664	163-243
Ge.met1	78222298	A1W1MR2D1B2	F7b	303-678	162-247
Ge.met2	78223514	Y3A2-W3-W2-R3B3-Y2	F10	301-681	176-252
Ge.met3	78223630	Y5--Y4A3--MMMW4R4D2B4	F8	320-691	154-239
Ge.met4	78223906	W5R6W6MA4B6	Alt	186-633	
Ge.met5	78224457	W7A5R8-B8-R9	F4b	211-593	
Ge.sul1	39995405	W1A1R2-B1-R1	F4b	213-596	
Ge.sul2	39996392	Y2M-Y3A2---M---MW4MW5-MM	F8	319-690	153-238
Ge.sul3	39997320	Y6A3-W7-W6--R4B3-Y5	F10	303-683	176-252
Ge.sul4	39998289	R5MW10Y7A4CD3	F1a	163-537	
Gl.oxy	58039985	M-Y1AWRBY2	F7a	290-666	170-255
Ha.che1	83643355	M----Y1A1W1MMW2-R1D1B1	F7a	299-680	161-248
Ha.che2	83643436	Y3Y2W4MR2A2B2W3	Tfp	2121-2578	
Ha.che3	83646436	Y5-A3-MW5R3D2B3	F8	324-690	152-238
Ha.che4	83646565	W7R4W6MA4B4	Alt	214-665	
Ha.che5	83647836	Y6ZA5B5---W9W8	F6	413-789	
Ha.che6	83648472	Y7Y8W10-MA6-W11	Tfp	558-1009	
Ha.mar	55378897	w2BAR	F1a	290-662	168-246
Halob	15790089	W2YBAC2C1DR	F1a	297-668	168-245
He.aci	109947052	V1AW	F3	268-651	
He.hep	32266171	V1AW	F3	246-629	
He.pyl	15612054	V3AW	F3	276-659	
Id.loi	56460222	YZAB-W2W1	F6	331-709	
Janna	89055332	db2-Y2AWR2Y1	F7a	342-720	176-261
La.int	94987586	B-R-W1Y2AA	F4a	339-724	
La.int-N**	94987585	B-R-W1Y2AA	F4a		
Le.int1	24213951	W1A1/CB1Y1	F1a	307-689	198-276
Le.int2	24215125	Y2-A2MW3D1B3	F8	324-693	161-247
Li.inn	16799775	R-----V-YA	F1b	243-618	137-217
Li.mon	16802734	R-----V-YA	F1b	243-618	137-217
Ma.mag1	83309425	A1W1Y1B1R1	F5	208-586	
Ma.mag2	83312103	MB4R4A2--W4	Alt	164-558	
Me.ace1	20088913	w1m--Y1B1A1R1C1D1	F1a	259-634	
Me.ace2	20091884	MW2-Y2B3A2C2D2R3	F1a	502-875	191-268
Me.bar	73668520	wm-YB1AR1D	F1a	222-597	
Me.bur	91772414	MW-YB1ACDR1	F1a	515-890	276-353
Me.cap	53805035	W2R2W1AB2	Alt	190-601	

Table A.2 (continued)

ID	GI	Gene Neighborhood	Class	Range	P2
Me.fla1	91775599	Y1Y2W1MA1	Tfp	1048-1499	
Me.fla2	91776288	Y4A2W3MMR1DB1Y3Z	F7b	393-769	163-229
Me.hun1	88601429	Y1B1A1/C1Dc2	F1a	403-773	172-252, 283-363
Me.hun2	88601797	m--W3MA2	Alt	212-657	
Me.hun3	88602280	W7R3W6MA3B4	Alt	182-617	
Me.lot	13488379	MW2RW1AB	Alt	179-605	
Me.mar	45358490	WBADMrc1c2y	F1a	546-920	325-404, 430-511
Me.maz1	21226430	MW1-Y1B1A1C1D1R1	F1a	501-874	195-272
Me.maz2	21227427	w2m-Y2B2A2R2C2D2	F1a	295-670	
Mo.the	83589594	MWAB	F1a	288-671	165-245
My.xan1	108756845	Y7A1W9W4--R8B1Y5	F10	466-850	174-250
My.xan2	108759279	Y8W2R5W6A2B5M	Alt	176-576	
My.xan3	108759462	W5R2W3MA3B3---y1--b2r6-mw1a7	Alt	184-585	
My.xan4	108760926	W14W11MA4B7R3	Alt	217-639	
My.xan5	108761228	W7-MMA5-B6R4	Alt	241-653	
My.xan6	108761624	W10R7MY4W8A6	Alt	282-709	
My.xan7	108763058	A7W1M-R6B2--Y1---b3a3mw3r2w5		282-660	
My.xan8	108763583	M-W13Y3A8C	F1a	477-854	368-446
Na.pha	76801730	mBAR	F1a	363-757	180-258
Ni.eur1	30249232	MA1	Tfp	1113-1564	
Ni.eur2	30249818	A2W3MM-RDB	F7b	349-725	186-254
Ni.ham	92118837	AWY2B2R2	F5	238-617	
Ni.mul	82701467	W1RW2MAB	Alt	185-630	
Ni.oce	77163664	Y2Y1W2MR1AB1W1	Tfp	1151-1606	
Ni.win	75674718	AWY1BR-----M	F5	216-595	
Nosto1	17228421	Y2Y1W1MA1	Tfp	1345-1820	
Nosto2	17228563	Y4Y3W2MA2	Tfp	522-998	
Nosto3	17229653	Y6Y5W3MA3	Tfp	338-807	
Oc.ihe	23099998	AW2	F1a	300-676	171-251
Pe.car1	77917649	MW1A1	F1a	231-593	
Pe.car2	77918801	A2W4R1BD---Y1X2X3	F7a	497-874	
Pe.car3	77919009	A3--Y2		264-642	
Ph.lum	37525780	AWMMRBYZ	F7b	286-662	168-234
Ph.pro1	54307969	MY2-A1-MR1B1	F8	358-726	164-250
Ph.pro2	54308135	Y3ZA2B2-W1W2	F6	362-739	
Polar1	91787038	Y1Y2W1MA1	Tfp	1451-1903	
Polar2	91788331	W2RW3--MMMA2B	Alt	185-630	
Ps.aer1	15595376	MY1A1W1MR1DB1	F7a	248-625	

Table A.2 (continued)

ID	GI	Gene Neighborhood	Class	Range	P2
Ps.aer2	15595610	Y2Y3W2MR2A2B2W3	Tfp	1854-2317	
Ps.aer3	15596655	Y4ZA3B3---W4W5	F6	370-745	
Ps.aer4	15598899	MW7R4W6A4B4	Alt	192-617	
Ps.arc	71066368	Y2Y1WMRA	Tfp	1675-2134	
Ps.atl	109899332	Y1ZAB-W2W1	F6	360-738	
Ps.cry	93006920	Y3Y2WMRAB	Tfp	1690-2149	
Ps.ent1	104780455	MW1R2W2A1B1	Alt	184-609	
Ps.ent2	104782798	Y1ZA2B3---W4W3	F6	351-726	
Ps.ent3	104783969	Y3Y2W6MA3W5	Tfp	1153-1608	
Ps.flu1	70728515	MW1R1W2A1B1	Alt	199-624	
Ps.flu2	70729060	Y1ZA2B2---W3W4	F6	375-750	
Ps.flu3	70733108	Y3Y2W6MA3W5	Tfp	1363-1820	
Ps.hal	77359761	Y1ZAB---W1W2	F6	342-718	
Ps.put1	26988225	MW1R1W2A1B1	Alt	185-610	
Ps.put2	26991028	Y1ZA2B3---W4W3	F6	363-738	
Ps.put3	26991665	Y3Y2W6MA3W5	Tfp	1041-1496	
Ps.syr1	28868133	MY1-A1MW1R1DB1	F8	306-677	157-243
Ps.syr2	28868704	MW2R2W3A2B2	Alt	215-641	
Ps.syr3	28869186	Y2ZA3B3---W4W5	F6	364-739	
Ps.syr4	28872144	Y4Y3W8MA4W7	Tfp	1384-1840	
Py.aby	14521752	wmRYBAC2C1DM	F1a	392-765	175-256, 270-352
Py.hor	14590396	wm-RYBAC1C2DM	F1a	386-759	171-252, 268-350
Ra.eut1	73537484	MW2R1W1A1B1	Alt	186-611	
Ra.eut2	73539430	M-Y1W3MM-----A2W4R4DB4Y2Z	F7b	283-659	163-229
Ra.eut3	73542304	Y5Y4W5MA3	Tfp	1322-1772	
Ra.met1	94309618	Y1Y2W1MA1	Tfp	1357-1807	
Ra.met2	94312621	M-Y3W2MM-----A2W3R2DB1Y4Z	F7b	282-658	161-227
Ra.met3	94312897	MW5R3W4A3B2	Alt	190-616	
Ra.sol1	17545391	Y1Y2W1MA1	Tfp	1426-1877	
Ra.sol2	17549627	Y5A2W2MR1DBY4Z2	F7b	310-686	174-240
Rh.cen1*	1621285	AWYBR	F5	225-603	
Rh.cen2*	31322729	MWB-RYA	F9	492-872	173-255
Rh.cen3*	31322737	YWRWMAB	Alt	161-591	
Rh.etl1	86356290	M-Y1A1W1R1B1Y2D	F7a	363-741	185-270
Rh.etl2	86359113	Y4A2W4MMMW3R2B2	F8	297-668	152-237

Table A.2 (continued)

ID	GI	Gene Neighborhood	Class	Range	P2
Rh.fer1	89899378	Y1A1W2R1D1B1	F7a	347-723	187-258
Rh.fer2	89899709	MY2-A2MW3R2D2B2	F8	338-704	165-251
Rh.fer3	89900169	Y3Y4W6MA3	Tfp	1275-1727	
Rh.pal1	39933219	Y1A1W2W1MR1B1	F8	302-672	152-237
Rh.pal2	39934697	A2W3Y3B2R2	F5	257-636	
Rh.pal3	39934746	MA3W4R3	F5	257-636	
Rh.rub1	83591860	A1Y1B1R1	F5	249-639	
Rh.rub2	83592735	Y2A2W1MW2MMR2B2Dm	F8	307-678	149-234
Rh.rub3	83593662	MW3B4-R4Y3A3--Y4--M	F9	474-861	173-255
Rh.sph1	77462124	Y1A1W1W2R1B1	F8	268-639	143-226
Rh.sph2	77462994	Y4M-MD-Y3A2W3R3Y2	F7a	294-672	165-250
Rh.sph3	77463620	A3R4B3W4-MY5A4	F7a	14-393	
Rh.sph4***	77463613	A3R4B3W4-MY5A4	F7a		529-624
Sa.deg1	90020168	Y2Y1W1MA1	Tfp	424-879	
Sa.deg2	90021806	Y4ZA2B1---W3W2	F6	378-759	
Sa.deg3	90022749	mm--Y5A3W4MR3DB3	F7a	396-775	152-237
Sa.deg4	90023269	Y7Y6W6MR4A4B4W5	Tfp	1730-2185	
Sa.ent	16760870	AWMRBYZ	F7b	281-657	165-231
Sa.rub	83815305	WR2Y-AB1-M-M-M		117-519	
Sh.den1	91792704	Y1ZA1B1--W1W2	F6	347-725	
Sh.den2	91794644	MY2-A2MW3R2DB2	F8	343-716	166-252
Sh.fle	30063340	AWMMRBYZ	F7b	264-640	160-226
Sh.one1	24373681	M--Y1A1W1MR1D1B1	F7a	321-696	172-253
Sh.one2	24374719	Y4ZA2B3--W4W3	F6	371-749	
Sh.son	74311762	AWMMRBYZ	F7b	264-640	160-226
Si.mel1	15964392	M-Y1A1W1R1B1Y2D	F7a	361-739	180-265
Si.mel2	16263301	R2W4MA2B2	Alt	166-566	
Silic	99078184	d1b1m-Y2AW1R1Y1	F7a	311-689	155-240
Sp.ala	103487225	AWYBR	F5	163-543	
Sy.aci1	85858537	W1R1W2MA1B1	Alt	194-639	
Sy.aci2	85859040	m--w3---MR2D1-Y1-D2B2A2Y2X1	F7a	352-728	
Sy.elo1	56750541	W1MA1	Tfp	295-775	
Sy.elo2	56750690	Y3Y2W3MA2W2	Tfp	334-799	
Sy.the	51892675	W2W3ACDYBRX1	F1a	309-694	173-253
Synco1	86606795	Y3Y2W1---MA1	Tfp	319-781	
Synco2	86607346	Y4Y5Y6W2MA2	Tfp	1107-1592	
Syncy1	16329790	Y4Y3W2MA1	Tfp	307-778	
Syncy2	16331224	A2	Tfp	277-768	
Syncy2-N**	16331742	A2	Tfp		
Syncy3	16331986	Y6Y5W4MMA3W3	Tfp	765-1245	
Tb.den1	74317642	Y2A1W1M-MMMRDBY1Z	F7b	301-677	164-230
Tb.den2	74318566	Y4Y3W2MA2	Tfp	1355-1805	

Table A.2 (continued)

ID	GI	Gene Neighborhood	Class	Range	P2
Th.elo1	22297892	Y1Y2W1MA1	Tfp	793-1271	
Th.elo2	22298111	Y4Y3W3MA2W2	Tfp	329-787	
Th.elo3	22298565	Y6Y5W4MA3	Tfp	262-717	
Th.kod	57640570	wmRYBAAC1C2MD	F1a	1-223	
Th.kod-N**	57640569	wmRYBAAC1C2MD	F1a	410-548	174-255, 282-364
Th.mar	15643465	AW1Y	F1a	296-670	179-260
Th.ten1	20807518	MW2B1-R1Y1A1	F9	418-799	165-251
Th.ten2	20807864	B2A2W4CD	F1a	271-645	160-240
Tm.cru1	78485095	Y1ZA1W2--y2w3-RBM	F6	268-644	
Tm.cru2	78485953	DA2	F7a	299-680	157-243
Tm.den1	78777172	Y1W1M-A1R1-DB1-Z1M-M----mM	F7a	304-681	175-263
Tm.den2	78777727	V2A2W2	F3	261-644	
Tr.den	42526999	AR2/W1XY	F2	342-781	192-271
Tr.pal	15639354	AR1/W1XY	F2	353-792	199-278
Vi.cho1	15601844	Y1A1W2W1MR1DM	F7a	323-702	166-252
Vi.cho2	15641408	mm-MW3B1-R2Y3A2-M	F9	372-756	175-254
Vi.cho3	15642063	Y4ZA3B2-W5W4	F6	398-775	
Vi.fis	59712438	YZAB--W2W1	F6	350-726	
Vi.par	28899003	YZAB-W2W1	F6	357-734	
Vi.vul1	27365300	Y1ZA1B1-W1W2	F6	354-731	
Vi.vul2	27367549	Y2A2W4W3MR3DB3M	F7a	307-686	170-256
Wo.suc	34558367	V4AW3	F3	259-642	
Xa.axo1	21242647	W5-Y2A1M-MM-MMMMMMMR2DB2	F7a	283-658	148-233
Xa.axo2	21243592	A2MW2-R3B3	F8	290-656	153-239
Xa.axo3	21243825	Y6Y5W4MA3B4W3	Tfp	1824-2274	
Xa.axo4	77748623	Y3ZA4	F6	174-547	
Xa.cam1	21231333	W6-Y2A1M--MMMM-MM-M-W1-MR2DB2	F7a	282-657	148-233
Xa.cam2	21231351	Y3ZA2	F6	173-546	
Xa.cam3	21232131	A3MW3-R3B4	F8	294-660	156-242
Xa.cam4	21232352	Y6Y5W5MA4B5W4	Tfp	1747-2197	
Xa.ory1	58581093	A1MW1-R1B1	F8	294-660	153-239
Xa.ory2-N**	58581373	Y6Y2W2MA2A2B2W3	Tfp		
Xa.ory2	58581374	Y6Y2W2MA2A2B2W3	Tfp	64-514	
Xa.ory3	58582247	Y4ZA3	F6	174-547	
Xa.ory4	58582459	W4-Y5A4-MM-M-MMMM-W5----R2DB3	F7a	284-659	148-233
Xy.fas	15838546	Y2W2MABW1	Tfp	1158-1607	
Ye.pes	16121930	AW	F7b	330-706	182-248
Ye.pse	51596728	m-----AW--MMRBYZ	F7b	330-706	182-248
Zy.mob1	56550979	M-A1RB1DYW	F7a	386-759	206-291
Zy.mob2	56551776	M-A2-B2		178-550	

Table A.3 CheB data. ID, GI, Gene Neighborhoods, and Range are explained in Table A.2. The R pair is the cognate CheR protein in Table A.4 that was used in the CheBR concatenated alignment and phylogenetic analysis. The class corresponds to the groups identified in Figure 3.9. The TCS-n and TCS-f classes correspond to the CheB and CheR proteins associated with TCSs by gene neighborhood and gene fusion, respectively. A minimum evolution tree was built from the concatenated CheBR alignment in MEGA with pairwise deletions and the JTT distance matrix.

ID	GI	Gene Neighborhood	R pair	Class	Range
Ac.bac1	94968556	A1W1MY1B1R1	Ac.bac1	F5	161-344
Ac.bac2	94968797	MW2R2A2B2Y2X1	Ac.bac2	F2	170-353
Ag.tum	15887867	M-Y1ARBY2D-----M	Ag.tum	F7	160-342
An.deh1	86157028	r2b2a2mw3w2Y3--Y2-D1A1W1R1B1	An.deh1	F8	166-347
An.deh2	86157042	b1r1w1a1d1-y2--y3W2W3MA2B2R2	An.deh2	Alt	160-339
An.deh3	86157618	R3W4MA3B3	An.deh3	Alt	158-339
An.deh4	86157799	b5r5w6ma5y5--W5MB4d2r4a4	An.deh4	F8	168-350
An.deh5	86157809	A4R4D2b4mw5--Y5A5MW6R5B5	An.deh5	F8	178-360
An.deh6	86159152	A6-W9W8-R6B6Y7	An.deh6	F10	156-343
An.deh7	86159603	B7/R8	An.deh8	TCS-f	65-243
An.var1	75906539	B1/R1	An.var1	TCS-f	28-207
An.var2	75910978	R2B2	An.var2	TCS-n	5-184
Ar.ful	11498646	MW-YBACDR--M	Ar.ful	F1	158-341
Azoar1	56476627	B1/R2-B2	Azoar2	TCS-f	22-201
Azoar2	56476629	B1/R2-B2			11-190
Ba.cla	56964012	BCD	Ba.cla	F1	118-300
Ba.hal	15614998	BCD	Ba.hal	F1	160-343
Ba.lic	52080245	BAWCD	Ba.lic	F1	167-351
Ba.sub	16078705	BAWCD	Ba.sub	F1	167-351
Bd.bac1	42522175	A1W1D1/B1			332-514
Bd.bac2	42524820	w3-A2R2B2	Bd.bac2	F7	165-347
Bo.bro	33601529	Y1AWMRBY2Z	Bo.bro	F7	159-342
Bo.bur1	15594760	R2B1	Bo.bur2	F2	189-371
Bo.bur2	15594913	W2-A1B2-Y2			202-384
Bo.gar1	51598672	R2-B1	Bo.gar2	F2	189-371
Bo.gar2	51598819	W2-A1B2-Y2			202-384
Bo.par	33596130	Y1AWMRBY2Z	Bo.par	F7	159-342
Bo.per	33592176	AWMBYZ			159-342
Br.jap1	27377306	A2W2Y2B1R2	Br.jap2	F5	207-391
Br.jap2	27377460	Y3A3W3MW4MR3B2	Br.jap3	F8	171-353
Bu.cen1	107024389	Y2A1W1MR1DB1Y1Z	Bu.cen1	F7	167-349
Bu.cen2	107026658	B2			151-331
Bu.cen3	107026914	MW3R2-W2A2B3	Bu.cen2	Alt	152-332
Bu.mal1	53724317	Y4A2W4MR2DB1-Y3Z	Bu.mal2	F7	171-353
Bu.mal2	77358928	MW2R1W1A1B2	Bu.mal1	Alt	153-333
Bu.pse1	53720911	Y2A1W2MR1DB1Y1Z	Bu.pse1	F7	174-356
Bu.pse2	53722890	MW4R2W3A2B2	Bu.pse2	Alt	153-333
Bu.tha1	83716866	Y3A1W2MR1D1B2B1			175-357
Bu.tha2	83717224	Y3A1W2MR1D1B2B1	Bu.tha1	F7	211-393
Bu.tha3	83717477	MW1R2W3A2B3	Bu.tha2	Alt	153-333

Table A.3 (continued)

ID	GI	Gene Neighborhood	R pair	Class	Range
Bu.tha4	83720721	Y4A3W4MR3D2B4Y5Z	Bu.tha3	F7	171-353
Bu.xen1	91778586	B1/R1	Bu.xen1	TCS-f	1-158
Bu.xen2	91778880	R2B2	Bu.xen2	TCS-n	24-204
Bu.xen3	91780564	B3/R3	Bu.xen3	TCS-f	19-199
Bu.xen4	91785653	Y4AWMR4DB4Y3Z	Bu.xen4	F7	169-351
Burkh1	78062866	B1			151-331
Burkh2	78063150	MW2R1W1A1B2	Burkh1	Alt	152-332
Burkh3	78064840	Y3A2W3MR2DB3Y4Z	Burkh2	F7	170-352
Ca.cre1	16124691	M-M-Y1A1W1R1B1Y2D	Ca.cre1	F7	157-339
Ca.cre2	16124851	mzy3-MA2W2Y4B2R2	Ca.cre2	F5	158-342
Ca.hyd1	78043438	MW2A1B1	Ca.hyd2	F1	154-339
Ca.hyd2	78043599	DR2CY1----M-MY3-R1MW1B2Y2A2	Ca.hyd1	F9	166-352
Ca.jej	15792253	BR	Ca.jej	F3	3-180
Ch.chl	78188541	B/R	Ch.chl	TCS-f	63-242
Ch.sal	92114142	AWMRBMYZ	Ch.sal	F7	162-344
Ch.vio1	34496464	Y1-A1MW1MD1B1			175-356
Ch.vio2	34497035	B2/R1	Ch.vio1	TCS-f	78-257
Ch.vio3	34497961	MY2-A2MW2R2B3	Ch.vio2	F8	175-357
Ch.vio4	34498594	B4			60-243
Ch.vio5	34498891	a4zy5v3v2--Y4A3W3--MR3B5D2	Ch.vio3	F7	166-349
Cl.ace	15895490	W3DBR2A2CY3W2	Cl.ace2	F1	159-339
Cl.tet	28211385	W2DBRACYW1	Cl.tet	F1	71-250
Co.psy	71279726	Y2ZAB---W2W1	Co.psy	F6	191-375
De.aro1	71906371	MY1A1W1MMR1D1B1-V1V2Y2ZA2	De.aro1	F7	165-347
De.aro2	71906787	Y3M-A3MW3D2R2D3B2	De.aro2	F8	173-355
De.aro3	71907673	B3/R3	De.aro3	TCS-f	25-204
De.des1	78356614	B1-R2-W3Y3A1	De.des2	F4a	183-366
De.des2	78357147	MY5A2R3B2	De.des3	F8	171-353
De.haf	89895738	m--AW2B	De.haf	F1	204-388
De.psy1	51245643	R1B1	De.psy1	TCS-n	6-186
De.psy2	51246438	B2			207-387
De.psy3	51246496	AW2B3R2---x1x2y2	De.psy2	F5	175-360
De.vul1	46578865	B1/R1	De.vul1	TCS-f	30-210
De.vul2	46580007	Y2A1R2B2	De.vul2	F8	171-353
De.vul3	46580483	B3-R3-W4Y3A3	De.vul3	F4a	181-364
Er.car	50120627	m-----AWMRBYZ	Er.car	F7	159-341
Er.lit	85375077	AWYBR-M	Er.lit	F5	154-337
Es.col	15802295	AWMMRBYZ	Es.col	F7	158-340
Ge.kau	56419776	BAW1CD	Ge.kau	F1	162-346
Ge.met1	78222000	B1/R1	Ge.met1	TCS-f	36-215
Ge.met2	78222293	A1W1MR2D1B2	Ge.met2	F7	170-352
Ge.met3	78223507	Y3A2-W3-W2-R3B3-Y2	Ge.met3	F10	169-352
Ge.met4	78223621	Y5--Y4A3--MMMW4R4D2B4	Ge.met4	F8	171-353
Ge.met5	78223837	R5B5	Ge.met5	TCS-n	11-191
Ge.met6	78223907	W5R6W6MA4B6	Ge.met6	Alt	162-343

Table A.3 (continued)

ID	GI	Gene Neighborhood	R pair	Class	Range
Ge.met7	78224022	M-B7M			6-185
Ge.met8	78224460	W7A5R8-B8-R9	Ge.met8	F4b	170-355
Ge.sul1	39995402	W1A1R2-B1-R1	Ge.sul2	F4b	175-360
Ge.sul2	39996247	MMW3R3D2B2	Ge.sul3	F8	170-352
Ge.sul3	39997312	Y6A3-W7-W6--R4B3-Y5	Ge.sul4	F10	180-363
Gl.oxy	58039988	M-Y1AWRBY2	Gl.oxy	F7	156-341
Gl.vio1	37521423	B1/R1	Gl.vio1	TCS-f	21-200
Gl.vio2	37523131	B2/R3	Gl.vio3	TCS-f	20-198
Ha.che1	83643363	M----Y1A1W1MMW2-R1D1B1	Ha.che1	F7	173-355
Ha.che2	83643435	Y3Y2W4MR2A2B2W3	Ha.che2	Tfp	154-334
Ha.che3	83646430	Y5-A3-MW5R3D2B3	Ha.che3	F8	175-357
Ha.che4	83646564	W7R4W6MA4B4	Ha.che4	Alt	163-343
Ha.che5	83647835	Y6ZA5B5---W9W8	Ha.che5	F6	203-386
Ha.mar	55378896	w2BAR	Ha.mar	F1	179-363
Halob	15790090	W2YBAC2C1DR	Halob	F1	158-342
He.hep	32265955	RB	He.hep	F3	26-209
Id.loi	56460221	YZAB-W2W1	Id.loi	F6	196-380
Janna1	89055055	B1/R1	Janna1	TCS-f	21-201
Janna2	89055335	y1r2way2-B2D	Janna2	F7	155-335
La.int	94987579	B-R-W1Y2AA	La.int	F4a	171-354
Le.int1	24213952	W1A1/CB1Y1			168-351
Le.int2	24214444	R1B2	Le.int1	TCS-n	5-185
Le.int3	24215129	Y2-A2MW3D1B3			165-347
Ma.mag1	83309422	A1W1Y1B1R1	Ma.mag1	F5	193-377
Ma.mag2	83311065	M--W2R2W3MB2	Ma.mag2	Alt	167-346
Ma.mag3	83312031	y6-----B3/R3	Ma.mag3	TCS-f	7-187
Ma.mag4	83312101	MB4R4A2--W4	Ma.mag4	Alt	166-346
Ma.mag5	83312428	B5/R5	Ma.mag5	TCS-f	11-190
Ma.mag6	83312753	B6/R6	Ma.mag6	TCS-f	16-195
Ma.mag7	83312979	R7B7	Ma.mag7	TCS-n	6-186
Me.ace1	20088914	w1m--Y1B1A1R1C1D1	Me.ace1	F1	167-354
Me.ace2	20090837	B2/R2	Me.ace2	TCS-f	76-257
Me.ace3	20091885	MW2-Y2B3A2C2D2R3	Me.ace3	F1	157-341
Me.ace4	20092349	B4/R4	Me.ace4	TCS-f	89-270
Me.bar1	73668521	wm-YB1AR1D	Me.bar1	F1	175-362
Me.bar2	73669676	B2/R2	Me.bar2	TCS-f	11-192
Me.bur1	91772413	MW-YB1ACDR1	Me.bur1	F1	159-343
Me.bur2	91772449	B2/R2	Me.bur2	TCS-f	5-183
Me.cap1	53804418	B1/R1	Me.cap1	TCS-f	1-171
Me.cap2	53805034	W2R2W1AB2	Me.cap2	Alt	158-338
Me.fla1	91776282	Y4A2W3MMR1DB1Y3Z	Me.fla1	F7	162-344
Me.fla2	91776939	R2B2	Me.fla2	TCS-n	16-196
Me.hun1	88601428	Y1B1A1/C1Dc2	Me.hun2	F1	147-336
Me.hun2	88602178	B2			3-182
Me.hun3	88602243	B3m			3-182
Me.hun4	88602279	W7R3W6MA3B4	Me.hun3	Alt	164-344

Table A.3 (continued)

ID	GI	Gene Neighborhood	R pair	Class	Range
Me.lot	13488378	MW2RW1AB	Me.lot	Alt	154-334
Me.mar	45358489	WBADMrc1c2y	Me.mar	F1	176-359
Me.maz1	21226431	MW1-Y1B1A1C1D1R1	Me.maz1	F1	165-349
Me.maz2	21227428	w2m-Y2B2A2R2C2D2	Me.maz2	F1	178-365
Mo.the	83589595	MWAB	Mo.the	F1	235-419
My.avi1	41409330	B1r-b2	My.avi	TCS-n	21-197
My.avi2	41409333	B2-Rb1			17-196
My.xan1	108758247	Y7A1W9W4--R8B1Y5	My.xan8	F10	157-340
My.xan2	108758271	A7W1M-R6B2--Y1---b3a3mw3r2w5	My.xan6	Alt	163-346
My.xan3	108760148	W5R2W3MA3B3---y1--b2r6-mw1a7	My.xan2	Alt	166-346
My.xan4	108761052	R9B4	My.xan9	TCS-n	6-182
My.xan5	108762177	Y8W2R5W6A2B5M	My.xan5	Alt	160-338
My.xan6	108762780	W7-MMA5-B6R4	My.xan4	Alt	162-342
My.xan7	108763031	W14W11MA4B7R3	My.xan3	Alt	150-330
Na.pha	76801731	mBAR	Na.pha	F1	167-351
Ni.eur	30249811	A2W3MM-RDB	Ni.eur	F7	165-347
Ni.ham1	92109717	B1/R1	Ni.ham1	TCS-f	34-214
Ni.ham2	92118834	AWY2B2R2	Ni.ham2	F5	191-375
Ni.ham3	92119393	B3/R3	Ni.ham3	TCS-f	20-199
Ni.mul	82701468	W1RW2MAB	Ni.mul	Alt	164-344
Ni.oce1	77163663	Y2Y1W2MR1AB1W1	Ni.oce1	Tfp	138-318
Ni.oce2	77165081	B2/R2	Ni.oce2	TCS-f	13-191
Ni.win	75674721	AWY1BR-----M	Ni.win	F5	235-419
Nosto1	17229208	B1/R1	Nosto1	TCS-f	1-138
Nosto2	17229339	R2B2	Nosto2	TCS-n	5-184
Oc.ihe	23099033	BW1CD	Oc.ihe	F1	165-350
Pe.car	77918804	A2W4R1BD---Y1X2X3	Pe.car1	F7	165-347
Ph.lum	37525785	AWMMRBYZ	Ph.lum	F7	159-341
Ph.pro1	54307973	MY2-A1-MR1B1	Ph.pro1	F8	166-348
Ph.pro2	54308136	Y3ZA2B2-W1W2	Ph.pro2	F6	208-391
Polar	91788332	W2RW3--MMMA2B	Polar	Alt	171-351
Ps.aer1	15595371	MY1A1W1MR1DB1	Ps.aer1	F7	157-339
Ps.aer2	15595611	Y2Y3W2MR2A2B2W3	Ps.aer2	Tfp	151-331
Ps.aer3	15596656	Y4ZA3B3---W4W5	Ps.aer3	F6	186-366
Ps.aer4	15598898	MW7R4W6A4B4	Ps.aer4	Alt	153-333
Ps.atl	109899331	Y1ZAB-W2W1	Ps.atl	F6	192-377
Ps.cry	93006919	Y3Y2WMRAB	Ps.cry	Tfp	138-318
Ps.ent1	104780456	MW1R2W2A1B1	Ps.ent2	Alt	154-334
Ps.ent2	104782257	R3B2	Ps.ent3	TCS-n	6-186
Ps.ent3	104782797	Y1ZA2B3---W4W3	Ps.ent4	F6	193-373
Ps.flu1	70728516	MW1R1W2A1B1	Ps.flu1	Alt	153-333
Ps.flu2	70729061	Y1ZA2B2---W3W4	Ps.flu3	F6	191-371
Ps.flu3	70730613	R2B3	Ps.flu2	TCS-n	16-196
Ps.hal	77359762	Y1ZAB---W1W2	Ps.hal	F6	203-387
Ps.put1	26988226	MW1R1W2A1B1	Ps.put1	Alt	154-334
Ps.put2	26990464	R2B2	Ps.put2	TCS-n	6-186

Table A.3 (continued)

ID	GI	Gene Neighborhood	R pair	Class	Range
Ps.put3	26991027	Y1ZA2B3---W4W3	Ps.put3	F6	188-368
Ps.syr1	28868128	MY1-A1MW1R1DB1	Ps.syr1	F8	176-358
Ps.syr2	28868705	MW2R2W3A2B2	Ps.syr2	Alt	153-333
Ps.syr3	28869187	Y2ZA3B3---W4W5	Ps.syr3	F6	208-388
Ps.syr4	28869901	R4B4	Ps.syr4	TCS-n	19-199
Py.aby	14521753	wmRYBAC2C1DM	Py.aby	F1	167-350
Py.hor	14590395	wm-RYBAC1C2DM	Py.hor	F1	167-350
Ra.eut1	73537483	MW2R1W1A1B1	Ra.eut1	Alt	152-332
Ra.eut2	73538740	B2/R2	Ra.eut2	TCS-f	47-226
Ra.eut3	73539014	B3/R3	Ra.eut3	TCS-f	15-196
Ra.eut4	73539434	M-Y1W3MM-----A2W4R4DB4Y2Z	Ra.eut4	F7	163-345
Ra.eut5	73540313	B5			5-183
Ra.met1	94312625	M-Y3W2MM-----A2W3R2DB1Y4Z	Ra.met2	F7	164-346
Ra.met2	94312896	MW5R3W4A3B2	Ra.met3	Alt	152-332
Ra.sol	17549622	Y5A2W2MR1DBY4Z2	Ra.sol1	F7	189-371
Rh.bal1	32473151	B1/R1	Rh.bal1	TCS-f	32-211
Rh.bal2	32476695	B2/R2	Rh.bal2	TCS-f	12-191
Rh.etl1	86356293	M-Y1A1W1R1B1Y2D	Rh.etl1	F7	156-338
Rh.etl2	86359106	Y4A2W4MMMW3R2B2	Rh.etl2	F8	179-361
Rh.etl3	86360808	B3B4/R3			9-188
Rh.etl4	86360809	B3B4/R3	Rh.etl3	TCS-f	28-206
Rh.fer1	89899382	Y1A1W2R1D1B1	Rh.fer1	F7	181-363
Rh.fer2	89899714	MY2-A2MW3R2D2B2	Rh.fer2	F8	166-348
Rh.fer3	89901132	m---MB3/R3-B4	Rh.fer3	TCS-f	24-205
Rh.fer4	89901134	m---MB3/R3-B4			1-146
Rh.fer5	89901784	B5/R4	Rh.fer4	TCS-f	15-194
Rh.pal1	39933214	Y1A1W2W1MR1B1	Rh.pal1	F8	171-353
Rh.pal2	39934700	A2W3Y3B2R2	Rh.pal2	F5	201-385
Rh.pal3	39936379	B3/R4	Rh.pal4	TCS-f	25-205
Rh.rub1	83591862	A1Y1B1R1	Rh.rub1	F5	198-383
Rh.rub2	83592742	Y2A2W1MW2MMR2B2Dm	Rh.rub2	F8	178-360
Rh.rub3	83592836	B3/R3	Rh.rub3	TCS-f	41-220
Rh.rub4	83593658	MW3B4-R4Y3A3--Y4--M	Rh.rub4	F9	172-356
Rh.sph1	77462128	Y1A1W1W2R1B1	Rh.sph1	F8	177-359
Rh.sph2	77462782	B2/R2	Rh.sph2	TCS-f	18-196
Rh.sph3	77463618	A3R4B3W4-MY5A4	Rh.sph4	F7	168-351
Sa.deg1	90021805	Y4ZA2B1---W3W2	Sa.deg1	F6	183-364
Sa.deg2	90022688	B2/R2	Sa.deg2	TCS-f	14-193
Sa.deg3	90022744	mm--Y5A3W4MR3DB3	Sa.deg3	F7	163-345
Sa.deg4	90023268	Y7Y6W6MR4A4B4W5	Sa.deg4	Tfp	120-301
Sa.ent	16760866	AWMRBYZ	Sa.ent	F7	158-340
Sa.rub1	83814685	WR2Y-AB1-M-M-M			214-397
Sa.rub2	83815839	B2/R1	Sa.rub1	TCS-f	95-273
Sh.boy	82543637	RBYZ	Sh.boy	F7	158-303
Sh.den1	91792705	Y1ZA1B1--W1W2	Sh.den1	F6	201-384
Sh.den2	91794649	MY2-A2MW3R2DB2	Sh.den2	F8	171-353

Table A.3 (continued)

ID	GI	Gene Neighborhood	R pair	Class	Range
Sh.fle	30063335	AWMMRBYZ	Sh.fle	F7	158-340
Sh.one1	24373686	M--Y1A1W1MR1D1B1	Sh.one1	F7	162-344
Sh.one2	24373874	Y2--MW2R2D2B2	Sh.one2	F8	171-353
Sh.one3	24374718	Y4ZA2B3--W4W3	Sh.one3	F6	187-370
Sh.son	74311767	AWMMRBYZ	Sh.son	F7	158-340
Si.mel1	15964395	M-Y1A1W1R1B1Y2D	Si.mel1	F7	158-340
Si.mel2	16263302	R2W4MA2B2	Si.mel2	Alt	161-341
Si.mel3	16264243	b4/r3B3			58-237
Si.mel4	16264244	b3B4/R3	Si.mel3	TCS-f	9-187
Silic1	99078188	y1r1w1ay2-MB1D1	Silic1	F7	127-308
Silic2	99080301	B2/R3	Silic3	TCS-f	9-192
Silic3	99080869	MB3/R4	Silic4	TCS-f	29-233
Silic4	99082067	B4/R5	Silic5	TCS-f	9-190
So.glo	85060043	WBZ			153-296
Sp.ala	103487222	AWYBR	Sp.ala	F5	176-359
Sy.aci1	85858538	W1R1W2MA1B1	Sy.aci1	Alt	172-352
Sy.aci2	85859039	m--w3---MR2D1-Y1-D2B2A2Y2X1	Sy.aci2	F7	160-342
Sy.the	51892679	W2W3ACDYBRX1	Sy.the	F1	178-358
Tb.den	74317633	Y2A1W1M-MMMRDBY1Z	Tb.den	F7	163-345
Th.kod	57640568	wmRYBAAC1C2MD	Th.kod	F1	166-349
Th.mar	15643174	B	Th.mar	F1	160-343
Th.ten1	20807514	MW2B1-R1Y1A1	Th.ten1	F9	171-357
Th.ten2	20807865	B2A2W4CD	Th.ten2	F1	6-187
Tm.cru	78485103	Y1ZA1W2--y2w3-RBM	Tm.cru	F7	191-373
Tm.den1	78777176	Y1W1M-A1R1-DB1-Z1M-M----mM	Tm.den1	F7	165-347
Tm.den2	78778133	B2R2	Tm.den2	F3	10-191
Tr.den	42526164	R1B	Tr.den1	F2	184-367
Tr.pal	15639619	R2B	Tr.pal2	F2	213-396
Vi.cho1	15641412	mm-MW3B1-R2Y3A2-M	Vi.cho2	F9	161-345
Vi.cho2	15642062	Y4ZA3B2-W5W4	Vi.cho3	F6	190-373
Vi.fis	59712437	YZAB--W2W1	Vi.fis	F6	189-372
Vi.par	28899002	YZAB-W2W1	Vi.par	F6	182-365
Vi.vul1	27365301	Y1ZA1B1-W1W2	Vi.vul1	F6	189-372
Vi.vul2	27366818	M----R2B2	Vi.vul2	TCS-n	10-190
Vi.vul3	27367543	Y2A2W4W3MR3DB3M	Vi.vul3	F7	159-342
Wo.suc	34557580	m--RB--m	Wo.suc	F3	35-217
Xa.axo1	21242033	R1B1	Xa.axo1	TCS-n	12-192
Xa.axo2	21242632	W5-Y2A1M-MM-MMMMMMMR2DB2	Xa.axo2	F7	171-353
Xa.axo3	21243597	A2MW2-R3B3	Xa.axo3	F8	168-350
Xa.axo4	21243824	Y6Y5W4MA3B4W3			225-392
Xa.axo5	77748740	B5/R4	Xa.axo4	TCS-f	23-202
Xa.cam1	21230640	R1B1	Xa.cam1	TCS-n	13-193
Xa.cam2	21231314	W6-Y2A1M--MMMM-MM-M-W1-MR2DB2	Xa.cam2	F7	171-353
Xa.cam3	21231469	B3			9-188
Xa.cam4	21232136	A3MW3-R3B4	Xa.cam3	F8	168-350
Xa.cam5	21232351	Y6Y5W5MA4B5W4			223-390

Table A.3 (continued)

ID	GI	Gene Neighborhood	R pair	Class	Range
Xa.cam6	77747953	B6/R4	Xa.cam4	TCS-f	22-201
Xa.cam7	77747904	B7			14-193
Xa.ory1	58581088	A1MW1-R1B1	Xa.ory1	F8	195-377
Xa.ory2	58581375	Y6Y2W2MA2A2B2W3			244-411
Xa.ory3	58582482	W4-Y5A4-MM-M-MMMM-W5----R2DB3	Xa.ory2	F7	171-353
Xa.ory4	77760529	B4/R3	Xa.ory3	TCS-f	23-202
Xy.fas	15838545	Y2W2MABW1			214-384
Ye.pes	16121941	MRBYZ	Ye.pes	F7	158-340
Ye.pse	51596721	m-----AW--MMRBYZ	Ye.pse	F7	158-340
Zy.mob1	56550977	M-A1RB1DYW	Zy.mob	F7	167-350
Zy.mob2	56551774	M-A2-B2		Class	156-338

Table A.4 CheR data. ID, GI, Gene Neighborhood, and Range are explained in Table A.2. Class is explained in Table A.3. The B pair is the cognate CheR protein from Table A.3 used in the CheBR concatenated alignement. A minimum evolution tree was built from the concatenated CheBR alignment in MEGA with pairwise deletions and the JTT distance matrix.

ID	GI	Gene Neighborhood	B pair*	Class	Range
Ac.bac1	94968557	A1W1MY1B1R1	Ac.bac1	F5	10-269
Ac.bac2	94968799	MW2R2A2B2Y2X1	Ac.bac2	F2	17-294
Ag.tum	15887866	M-Y1ARBY2D-----M	Ag.tum	F7	31-293
An.deh1	86157029	r2b2a2mw3w2Y3--Y2-D1A1W1R1B1	An.deh1	F8	31-291
An.deh2	86157043	b1r1w1a1d1-y2--y3W2W3MA2B2R2	An.deh2	Alt	11-272
An.deh3	86157622	R3W4MA3B3	An.deh3	Alt	17-275
An.deh4	86157797	A4R4D2b4mw5--Y5A5MW6R5B5	An.deh4	F8	8-264
An.deh5	86157808	A4R4D2b4mw5--Y5A5MW6R5B5	An.deh5	F8	36-299
An.deh6	86159153	A6-W9W8-R6B6Y7	An.deh6	F10	18-278
An.deh7	86159363	R7			41-295
An.deh8	86159603	B7/R8	An.deh7	TCS-f	265-518
An.deh9	86160370	W10R9			13-263
An.var1	75906539	B1/R1	An.var1	TCS-f	234-489
An.var2	75910977	R2B2	An.var2	TCS-n	15-271
Ar.ful	11498642	MW-YBACDR--M	Ar.ful	F1	8-263
Azoar1	56475845	R1			18-274
Azoar2	56476627	B1/R2-B2	Azoar1	TCS-f	225-480
Ba.ant1	49183953	R1			24-280
Ba.ant2	49184562	R2---a-y			10-258
Ba.cer1	30019162	R1			24-280
Ba.cer2	30019779	R2---ay			15-263
Ba.cla	56963656	R	Ba.cla	F1	10-258
Ba.hal	15614218	R	Ba.hal	F1	8-255
Ba.lic	52080781	R	Ba.lic	F1	7-253
Ba.sub	50812262	R	Ba.sub	F1	7-254
Ba.thu1	49480179	R1			24-280
Ba.thu2	49480948	R2---ay			15-263
Bd.bac1	42524239	MR1D2W2			15-271
Bd.bac2	42524821	w3-A2R2B2	Bd.bac2	F7	28-288
Bo.bro	33601528	Y1AWMRBY2Z	Bo.bro	F7	25-279
Bo.bur1	15594386	R1			21-282
Bo.bur2	15594759	R2B1	Bo.bur1	F2	11-265
Bo.bur3	15595015	A2W3/R3XY3			201-463
Bo.gar1	51598303	R1			21-282
Bo.gar2	51598670	R2-B1	Bo.gar1	F2	11-265
Bo.gar3	51598925	A2W3/R3XY3			201-464
Bo.par	33596129	Y1AWMRBY2Z	Bo.par	F7	25-279
Br.jap1	27375501	A1W1Y1R1			8-266
Br.jap2	27377307	A2W2Y2B1R2	Br.jap1	F5	8-266
Br.jap3	27377459	Y3A3W3MW4MR3B2	Br.jap2	F8	19-284
Bu.cen1	107024391	Y2A1W1MR1DB1Y1Z	Bu.cen1	F7	35-290
Bu.cen2	107026918	MW3R2-W2A2B3	Bu.cen3	Alt	7-275

Table A.4 (continued)

ID	GI	Gene Neighborhood	B pair*	Class	Range
Bu.mal1	53716490	MW2R1W1A1B2	Bu.mal2	Alt	7-278
Bu.mal2	53724319	Y4A2W4MR2DB1-Y3Z	Bu.mal1	F7	34-289
Bu.pse1	53720913	Y2A1W2MR1DB1Y1Z	Bu.pse1	F7	35-290
Bu.pse2	53722893	MW4R2W3A2B2	Bu.pse2	Alt	7-278
Bu.tha1	83716910	Y3A1W2MR1D1B2B1	Bu.tha2	F7	15-269
Bu.tha2	83717558	MW1R2W3A2B3	Bu.tha3	Alt	7-279
Bu.tha3	83718451	Y4A3W4MR3D2B4Y5Z	Bu.tha4	F7	34-289
Bu.xen1	91778586	B1/R1	Bu.xen1	TCS-f	181-434
Bu.xen2	91778881	R2B2	Bu.xen2	TCS-n	21-277
Bu.xen3	91780564	B3/R3	Bu.xen3	TCS-f	222-475
Bu.xen4	91785655	Y4AWMR4DB4Y3Z	Bu.xen4	F7	34-289
Burkh1	78063153	MW2R1W1A1B2	Burkh2	Alt	7-275
Burkh2	78064838	Y3A2W3MR2DB3Y4Z	Burkh3	F7	35-290
Ca.cre1	16124690	M-M-Y1A1W1R1B1Y2D	Ca.cre1	F7	29-290
Ca.cre2	16124852	mzy3-MA2W2Y4B2R2	Ca.cre2	F5	8-263
Ca.cre3	16127702	Y5R3			8-266
Ca.hyd1	78043366	DR2CY1----M-MY3-R1MW1B2Y2A2	Ca.hyd2	F9	8-266
Ca.hyd2	78043373	DR2CY1----M-MY3-R1MW1B2Y2A2	Ca.hyd1	F1	6-251
Ca.jej	15792252	BR	Ca.jej	F3	14-259
Ch.chl	78188541	B/R	Ch.chl	TCS-f	266-522
Ch.sal	92114143	AWMRBMYZ	Ch.sal	F7	34-288
Ch.vio1	34497035	B2/R1	Ch.vio2	TCS-f	272-528
Ch.vio2	34497962	MY2-A2MW2R2B3	Ch.vio3	F8	15-272
Ch.vio3	34498892	a4zy5v3v2--Y4A3W3--MR3B5D2	Ch.vio5	F7	30-288
Ch.vio4	34499148	R4			27-282
Cl.ace1	15893417	Y1A1W1MR1Y2			11-270
Cl.ace2	15895489	W3DBR2A2CY3W2	Cl.ace	F1	6-253
Cl.tet	28211384	W2DBRACYW1	Cl.tet	F1	8-255
Co.psy	71280037	V-R	Co.psy	F6	13-273
De.aro1	71906369	MY1A1W1MMR1D1B1-V1V2Y2ZA2	De.aro1	F7	41-296
De.aro2	71906785	Y3M-A3MW3D2R2D3B2	De.aro2	F8	14-269
De.aro3	71907673	B3/R3	De.aro3	TCS-f	231-487
De.des1	78356245	R1-X2			59-323
De.des2	78356616	B1-R2-W3Y3A1	De.des1	F4a	24-288
De.des3	78357148	MY5A2R3B2	De.des2	F8	20-282
De.haf	89895782	R	De.haf	F1	12-259
De.psy1	51245644	R1B1	De.psy1	TCS-n	45-301
De.psy2	51246497	AW2B3R2---x1x2y2	De.psy3	F5	11-275
De.vul1	46578865	B1/R1	De.vul1	TCS-f	239-494
De.vul2	46580006	Y2A1R2B2	De.vul2	F8	8-270
De.vul3	46580481	B3-R3-W4Y3A3	De.vul3	F4a	24-288
Er.car	50120626	m-----AWMRBYZ	Er.car	F7	32-286
Er.lit	85375076	AWYBR-M	Er.lit	F5	10-269
Es.col	15802296	AWMMRBYZ	Es.col	F7	29-283
Ge.kau	56420743	R	Ge.kau	F1	7-254

Table A.4 (continued)

ID	GI	Gene Neighborhood	B pair*	Class	Range
Ge.met1	78222000	B1/R1	Ge.met1	TCS-f	239-495
Ge.met2	78222295	A1W1MR2D1B2	Ge.met2	F7	33-288
Ge.met3	78223508	Y3A2-W3-W2-R3B3-Y2	Ge.met3	F10	17-279
Ge.met4	78223623	Y5--Y4A3--MMMW4R4D2B4	Ge.met4	F8	22-287
Ge.met5	78223838	R5B5	Ge.met5	TCS-n	13-269
Ge.met6	78223903	W5R6W6MA4B6	Ge.met6	Alt	12-278
Ge.met7	78224403	R7Y7CD3			26-282
Ge.met8	78224458	W7A5R8-B8-R9	Ge.met8	F4b	10-275
Ge.met9	78224462	W7A5R8-B8-R9			91-346
Ge.sul1	39995400	W1A1R2-B1-R1			83-336
Ge.sul2	39995404	W1A1R2-B1-R1	Ge.sul1	F4b	10-275
Ge.sul3	39996245	MMW3R3D2B2	Ge.sul2	F8	26-290
Ge.sul4	39997313	Y6A3-W7-W6--R4B3-Y5	Ge.sul3	F10	17-279
Ge.sul5	39998285	R5MW10Y7A4CD3			42-298
Gl.oxy	58039987	M-Y1AWRBY2	Gl.oxy	F7	19-277
Gl.vio1	37521423	B1/R1	Gl.vio1	TCS-f	226-480
Gl.vio2	37522899	R2			44-288
Gl.vio3	37523131	B2/R3	Gl.vio2	TCS-f	220-474
Ha.che1	83643361	M----Y1A1W1MMW2-R1D1B1	Ha.che1	F7	23-280
Ha.che2	83643437	Y3Y2W4MR2A2B2W3	Ha.che2	Tfp	28-289
Ha.che3	83646432	Y5-A3-MW5R3D2B3	Ha.che3	F8	26-278
Ha.che4	83646568	W7R4W6MA4B4	Ha.che4	Alt	24-291
Ha.che5	83647405	V1-R5	Ha.che5	F6	1-234
Ha.mar	55378898	w2BAR	Ha.mar	F1	12-265
Halob	15790085	W2YBAC2C1DR	Halob	F1	1-227
He.hep	32265954	RB	He.hep	F3	11-275
Id.loi	56460256	VR	Id.loi	F6	13-275
Janna1	89055055	B1/R1	Janna1	TCS-f	223-477
Janna2	89055330	db2-Y2AWR2Y1	Janna2	F7	25-286
La.int	94987581	B-R-W1Y2AA	La.int	F4a	23-287
Le.int1	24214443	R1B2	Le.int2	TCS-n	66-322
Le.int2	24214742	R2			31-299
Li.inn	16799766	R-----V-YA			12-260
Li.mon	16802725	R-----V-YA			12-260
Ma.mag1	83309421	A1W1Y1B1R1	Ma.mag1	F5	8-266
Ma.mag2	83311062	M--W2R2W3MB2	Ma.mag2	Alt	23-288
Ma.mag3	83312031	y6-----B3/R3	Ma.mag3	TCS-f	214-468
Ma.mag4	83312102	MB4R4A2--W4	Ma.mag4	Alt	8-267
Ma.mag5	83312428	B5/R5	Ma.mag5	TCS-f	212-465
Ma.mag6	83312753	B6/R6	Ma.mag6	TCS-f	214-468
Ma.mag7	83312978	R7B7	Ma.mag7	TCS-n	15-271
Me.ace1	20088912	w1m--Y1B1A1R1C1D1	Me.ace1	F1	117-377
Me.ace2	20090837	B2/R2	Me.ace2	TCS-f	278-533
Me.ace3	20091881	MW2-Y2B3A2C2D2R3	Me.ace3	F1	12-266
Me.ace4	20092349	B4/R4	Me.ace4	TCS-f	291-546
Me.bar1	73668519	wm-YB1AR1D	Me.bar1	F1	41-297

Table A.4 (continued)

ID	GI	Gene Neighborhood	B pair*	Class	Range
Me.bar2	73669676	B2/R2	Me.bar2	TCS-f	213-468
Me.bur1	91772417	MW-YB1ACDR1	Me.bur1	F1	12-266
Me.bur2	91772449	B2/R2	Me.bur2	TCS-f	207-460
Me.cap1	53804418	B1/R1	Me.cap1	TCS-f	186-440
Me.cap2	53805037	W2R2W1AB2	Me.cap2	Alt	8-262
Me.fla1	91776284	Y4A2W3MMR1DB1Y3Z	Me.fla1	F7	24-278
Me.fla2	91776938	R2B2	Me.fla2	TCS-n	25-281
Me.hun1	88601442	Y2-R1			1-198
Me.hun2	88602252	W5R2	Me.hun1	F1	26-278
Me.hun3	88602283	W7R3W6MA3B4	Me.hun4	Alt	12-278
Me.lot	13488381	MW2RW1AB	Me.lot	Alt	4-255
Me.mar	45358493	YC2C1Rmdabw	Me.mar	F1	9-263
Me.maz1	21226427	MW1-Y1B1A1C1D1R1	Me.maz1	F1	12-266
Me.maz2	21227426	w2m-Y2B2A2R2C2D2	Me.maz2	F1	39-295
Mo.the	83589653	DRCY	Mo.the	F1	6-253
My.avi	41409331	B2-Rb1	My.avi1	TCS-n	15-268
My.xan1	108756964	R1			81-340
My.xan2	108757252	W5R2W3MA3B3---y1--b2r6-mw1a7	My.xan3	Alt	13-270
My.xan3	108757834	W14W11MA4B7R3	My.xan7	Alt	11-271
My.xan4	108758057	W7-MMA5-B6R4	My.xan6	Alt	9-263
My.xan5	108759708	Y8W2R5W6A2B5M	My.xan5	Alt	7-264
My.xan6	108760379	A7W1M-R6B2--Y1---b3a3mw3r2w5	My.xan2	Alt	13-268
My.xan7	108760571	W10R7MY4W8A6			13-261
My.xan8	108761128	Y7A1W9W4--R8B1Y5	My.xan1	F10	18-279
My.xan9	108762503	R9B4	My.xan4	TCS-n	18-273
Na.pha	76801729	mBAR	Na.pha	F1	13-268
Ni.eur	30249813	A2W3MM-RDB	Ni.eur	F7	23-279
Ni.ham1	92109717	B1/R1	Ni.ham1	TCS-f	240-492
Ni.ham2	92118833	AWY2B2R2	Ni.ham2	F5	8-266
Ni.ham3	92119393	B3/R3	Ni.ham3	TCS-f	224-432
Ni.mul	82701464	W1RW2MAB	Ni.mul	Alt	12-279
Ni.oce1	77163665	Y2Y1W2MR1AB1W1	Ni.oce1	Tfp	16-276
Ni.oce2	77165081	B2/R2	Ni.oce2	TCS-f	217-470
Ni.win	75674722	AWY1BR-----M	Ni.win	F5	8-266
Nosto1	17229208	B1/R1	Nosto1	TCS-f	165-419
Nosto2	17229340	R2B2	Nosto2	TCS-n	15-271
Oc.ihe	23099241	R	Oc.ihe	F1	7-254
Pe.car1	77918803	A2W4R1BD---Y1X2X3	Pe.car	F7	28-290
Pe.car2	77920085	R2			8-263
Ph.lum	37525784	AWMMRBYZ	Ph.lum	F7	33-287
Ph.pro1	54307972	MY2-A1-MR1B1	Ph.pro1	F8	47-302
Ph.pro2	54308093	V2R2	Ph.pro2	F6	13-273
Polar	91788324	W2RW3--MMMA2B	Polar	Alt	12-278
Ps.aer1	15595373	MY1A1W1MR1DB1	Ps.aer1	F7	24-279
Ps.aer2	15595609	Y2Y3W2MR2A2B2W3	Ps.aer2	Tfp	22-283
Ps.aer3	15598544	VR3	Ps.aer3	F6	10-271

Table A.4 (continued)

ID	GI	Gene Neighborhood	B pair*	Class	Range
Ps.aer4	15598901	MW7R4W6A4B4	Ps.aer4	Alt	4-259
Ps.arc	71066369	Y2Y1WMRA			12-295
Ps.atl	109899406	VR	Ps.atl	F6	13-274
Ps.cry	93006921	Y3Y2WMRAB	Ps.cry	Tfp	41-319
Ps.ent1	104779523	R1			14-257
Ps.ent2	104780453	MW1R2W2A1B1	Ps.ent1	Alt	5-260
Ps.ent3	104782258	R3B2	Ps.ent2	TCS-n	13-269
Ps.ent4	104782848	V3R4	Ps.ent3	F6	10-272
Ps.flu1	70728513	MW1R1W2A1B1	Ps.flu1	Alt	6-262
Ps.flu2	70730614	R2B3	Ps.flu3	TCS-n	20-276
Ps.flu3	70731827	V2R3	Ps.flu2	F6	10-272
Ps.hal	77359717	VR	Ps.hal	F6	13-273
Ps.put1	26988223	MW1R1W2A1B1	Ps.put1	Alt	5-260
Ps.put2	26990465	R2B2	Ps.put2	TCS-n	13-269
Ps.put3	26991081	V3R3	Ps.put3	F6	10-272
Ps.syr1	28868130	MY1-A1MW1R1DB1	Ps.syr1	F8	14-268
Ps.syr2	28868702	MW2R2W3A2B2	Ps.syr2	Alt	7-261
Ps.syr3	28869132	V2R3	Ps.syr3	F6	15-277
Ps.syr4	28869900	R4B4	Ps.syr4	TCS-n	28-284
Py.aby	14521755	wmRYBAC2C1DM	Py.aby	F1	10-269
Py.hor	14590393	wm-RYBAC1C2DM	Py.hor	F1	10-269
Ra.eut1	73537486	MW2R1W1A1B1	Ra.eut1	Alt	7-260
Ra.eut2	73538740	B2/R2	Ra.eut2	TCS-f	257-512
Ra.eut3	73539014	B3/R3	Ra.eut3	TCS-f	220-474
Ra.eut4	73539432	M-Y1W3MM-----A2W4R4DB4Y2Z	Ra.eut4	F7	41-293
Ra.met1	94311884	R1			193-455
Ra.met2	94312623	M-Y3W2MM-----A2W3R2DB1Y4Z	Ra.met1	F7	44-296
Ra.met3	94312899	MW5R3W4A3B2	Ra.met2	Alt	10-267
Ra.sol1	17549624	Y5A2W2MR1DBY4Z2	Ra.sol	F7	26-281
Ra.sol2	17549866	R2			7-272
Rh.bal1	32473151	B1/R1	Rh.bal1	TCS-f	236-488
Rh.bal2	32476695	B2/R2	Rh.bal2	TCS-f	217-481
Rh.etl1	86356292	M-Y1A1W1R1B1Y2D	Rh.etl1	F7	1-259
Rh.etl2	86359107	Y4A2W4MMMW3R2B2	Rh.etl2	F8	23-284
Rh.etl3	86360809	B3B4/R3	Rh.etl4	TCS-f	232-485
Rh.fer1	89899380	Y1A1W2R1D1B1	Rh.fer1	F7	43-297
Rh.fer2	89899712	MY2-A2MW3R2D2B2	Rh.fer2	F8	11-265
Rh.fer3	89901132	m---MB3/R3-B4	Rh.fer3	TCS-f	230-486
Rh.fer4	89901784	B5/R4	Rh.fer5	TCS-f	220-476
Rh.pal1	39933215	Y1A1W2W1MR1B1	Rh.pal1	F8	15-280
Rh.pal2	39934701	A2W3Y3B2R2	Rh.pal2	F5	8-267
Rh.pal3	39934748	MA3W4R3			8-266
Rh.pal4	39936379	B3/R4	Rh.pal3	TCS-f	231-483
Rh.rub1	83591863	A1Y1B1R1	Rh.rub1	F5	8-266
Rh.rub2	83592741	Y2A2W1MW2MMR2B2Dm	Rh.rub2	F8	18-283
Rh.rub3	83592836	B3/R3	Rh.rub3	TCS-f	248-504

Table A.4 (continued)

ID	GI	Gene Neighborhood	B pair*	Class	Range
Rh.rub4	83593660	MW3B4-R4Y3A3--Y4--M	Rh.rub4	F9	10-271
Rh.sph1	77462127	Y1A1W1W2R1B1	Rh.sph1	F8	15-277
Rh.sph2	77462782	B2/R2	Rh.sph2	TCS-f	217-471
Rh.sph3	77462992	Y4M-MD-Y3A2W3R3Y2			24-284
Rh.sph4	77463619	A3R4B3W4-MY5A4	Rh.sph3	F7	11-267
Sa.deg1	90021859	VR1	Sa.deg1	F6	21-282
Sa.deg2	90022688	B2/R2	Sa.deg2	TCS-f	219-473
Sa.deg3	90022746	mm--Y5A3W4MR3DB3	Sa.deg3	F7	27-283
Sa.deg4	90023270	Y7Y6W6MR4A4B4W5	Sa.deg4	Tfp	18-279
Sa.ent	16760867	AWMRBYZ	Sa.ent	F7	29-283
Sa.rub1	83815839	B2/R1	Sa.rub2	TCS-f	299-553
Sa.rub2	83816103	WR2Y-AB1-M-M-M			18-280
Sh.boy	82543636	RBYZ	Sh.boy	F7	1-251
Sh.den1	91792660	V1R1	Sh.den1	F6	15-275
Sh.den2	91794647	MY2-A2MW3R2DB2	Sh.den2	F8	22-274
Sh.fle	30063336	AWMMRBYZ	Sh.fle	F7	29-283
Sh.one1	24373684	M--Y1A1W1MR1D1B1	Sh.one1	F7	14-266
Sh.one2	24373872	Y2--MW2R2D2B2	Sh.one2	F8	20-274
Sh.one3	24374762	V3R3	Sh.one3	F6	15-275
Sh.son	74311766	AWMMRBYZ	Sh.son	F7	29-283
Si.mel1	15964394	M-Y1A1W1R1B1Y2D	Si.mel1	F7	31-293
Si.mel2	16263298	R2W4MA2B2	Si.mel2	Alt	9-267
Si.mel3	16264244	b3B4/R3	Si.mel4	TCS-f	209-462
Silic1	99078182	d1b1m-Y2AW1R1Y1	Silic1	F7	23-282
Silic2	99078221	MR2D2W2			15-271
Silic3	99080301	B2/R3	Silic2	TCS-f	214-472
Silic4	99080869	MB3/R4	Silic3	TCS-f	257-514
Silic5	99082067	B4/R5	Silic4	TCS-f	212-466
Sp.ala	103487221	AWYBR	Sp.ala	F5	13-273
Sy.aci1	85858534	W1R1W2MA1B1	Sy.aci1	Alt	12-277
Sy.aci2	85859033	m--w3---MR2D1-Y1-D2B2A2Y2X1	Sy.aci2	F7	14-277
Sy.the	51892680	W2W3ACDYBRX1	Sy.the	F1	9-256
Tb.den	74317635	Y2A1W1M-MMMRDBY1Z	Tb.den	F7	24-278
Th.kod	57640566	wmRYBAAC1C2MD	Th.kod	F1	12-269
Th.mar	15643230	R	Th.mar	F1	32-279
Th.ten1	20807516	MW2B1-R1Y1A1	Th.ten1	F9	8-270
Th.ten2	20807802	R2	Th.ten2	F1	7-254
Tm.cru	78485102	Y1ZA1W2--y2w3-RBM	Tm.cru	F7	15-273
Tm.den1	78777173	Y1W1M-A1R1-DB1-Z1M-M----mM	Tm.den1	F7	15-280
Tm.den2	78778134	B2R2	Tm.den2	F3	29-259
Tr.den1	42526163	R1B	Tr.den	F2	12-267
Tr.den2	42527000	AW1/R2XY			194-442
Tr.pal1	15639355	AW1/R1XY			205-453
Tr.pal2	15639618	R2B	Tr.pal	F2	34-289
Vi.cho1	15601840	Y1A1W2W1MR1DM			18-275
Vi.cho2	15641410	mm-MW3B1-R2Y3A2-M	Vi.cho1	F9	13-273

Table A.4 (continued)

ID	GI	Gene Neighborhood	B pair*	Class	Range
Vi.cho3	15642200	V4R3	Vi.cho2	F6	13-272
Vi.fis	59712485	V1R	Vi.fis	F6	13-272
Vi.par	28897548	V1R	Vi.par	F6	13-272
Vi.vul1	27363709	V1R1	Vi.vul1	F6	13-272
Vi.vul2	27366817	M----R2B2	Vi.vul2	TCS-n	16-272
Vi.vul3	27367545	Y2A2W4W3MR3DB3M	Vi.vul3	F7	26-283
Wo.suc	34557579	m--RB--m	Wo.suc	F3	11-276
Xa.axo1	21242034	R1B1	Xa.axo1	TCS-n	13-269
Xa.axo2	21242634	W5-Y2A1M-MM-MMMMMMMR2DB2	Xa.axo2	F7	24-280
Xa.axo3	21243596	A2MW2-R3B3	Xa.axo3	F8	21-274
Xa.axo4	77748740	B5/R4	Xa.axo5	TCS-f	228-482
Xa.cam1	21230641	R1B1	Xa.cam1	TCS-n	13-269
Xa.cam2	21231316	W6-Y2A1M--MMMM-MM-M-W1-MR2DB2	Xa.cam2	F7	25-281
Xa.cam3	21232135	A3MW3-R3B4	Xa.cam4	F8	21-274
Xa.cam4	77747953	B6/R4	Xa.cam6	TCS-f	226-480
Xa.ory1	58581089	A1MW1-R1B1	Xa.ory1	F8	21-274
Xa.ory2	58582480	W4-Y5A4-MM-M-MMMM-W5----R2DB3	Xa.ory3	F7	24-280
Xa.ory3	77760529	B4/R3	Xa.ory4	TCS-f	228-482
Ye.pes	16121940	MRBYZ	Ye.pes	F7	32-286
Ye.pse	51596722	m-----AW--MMRBYZ	Ye.pse	F7	32-286
Zy.mob	56550978	M-A1RB1DYW	Zy.mob1	F7	27-287

Table A.5 CheD data. ID, GI, Gene Neighborhood, and Range are explained in Table A.2. A neighbor-joining tree was built from the CheD alignment in MEGA with pairwise deletions and the Poisson correction distance.

ID	GI	Gene Neighborhood	Range
Ag.tum	15887869	M-Y1ARBY2D-----M	9-152
An.deh1	86157032	r2b2a2mw3w2Y3--Y2-D1A1W1R1B1	28-171
An.deh2	86157798	A4R4D2b4mw5--Y5A5MW6R5B5	4-153
Ar.ful	11498643	MW-YBACDR--M	5-150
Ba.cla	56964010	BCD	5-152
Ba.hal	15614996	BCD	5-155
Ba.lic	52080249	BAWCD	11-161
Ba.sub	16078709	BAWCD	9-159
Bd.bac1	42522175	A1W1D1/B1	5-149
Bd.bac2	42524238	MR1D2W2	6-152
Bo.bro	33603425	D	23-169
Bo.bur	15594951	D	14-157
Bo.gar	51598859	D	14-157
Bo.par	33598482	D	23-169
Bo.per	33594688	D	23-169
Bu.cen	107024390	Y2A1W1MR1DB1Y1Z	1-148
Bu.mal	53724318	Y4A2W4MR2DB1-Y3Z	21-168
Bu.pse	53720912	Y2A1W2MR1DB1Y1Z	21-168
Bu.tha1	83716255	Y3A1W2MR1D1B2B1	16-162
Bu.tha2	83721660	Y4A3W4MR3D2B4Y5Z	21-168
Bu.xen	91785654	Y4AWMR4DB4Y3Z	22-169
Burkh	78064839	Y3A2W3MR2DB3Y4Z	21-168
Ca.cre	16124693	M-M-Y1A1W1R1B1Y2D	15-159
Ca.hyd	78044303	DR2CY1----M-MY3-R1MW1B2Y2A2	3-152
Ch.vio1	34496465	Y1-A1MW1MD1B1	30-178
Ch.vio2	34498890	a4zy5v3v2--Y4A3W3--MR3B5D2	22-168
Cl.ace	15895491	W3DBR2A2CY3W2	10-159
Cl.tet	28211386	W2DBRACYW1	9-158
De.aro1	71906370	MY1A1W1MMR1D1B1-V1V2Y2ZA2	24-169
De.aro2	71906784	Y3M-A3MW3D2R2D3B2	16-162
De.aro3	71906786	Y3M-A3MW3D2R2D3B2	11-158
De.des	78358252	D	34-186
De.haf	89895715	CD	5-153
De.vul	46581375	D	29-181
Ge.kau	56419780	BAW1CD	7-158
Ge.met1	78222294	A1W1MR2D1B2	22-164
Ge.met2	78223622	Y5--Y4A3--MMMW4R4D2B4	11-153
Ge.met3	78224400	R7Y7CD3	5-155
Ge.sul1	39995832	D1	31-184
Ge.sul2	39996246	MMW3R3D2B2	1-124
Ge.sul3	39998291	R5MW10Y7A4CD3	5-155
Ha.che1	83643362	M----Y1A1W1MMW2-R1D1B1	34-180
Ha.che2	83646431	Y5-A3-MW5R3D2B3	172-317
Ha.mar	55378886	YC3D	20-166

Table A.5 (continued)

ID	GI	Gene Neighborhood	Range
Halob	16554481	W2YBAC2C1DR	3-148
Janna	89055336	y1r2way2-B2D	1-140
Le.int1	24215128	Y2-A2MW3D1B3	11-163
Le.int2	24217186	D2	24-173
Me.ace1	20088910	w1m--Y1B1A1R1C1D1	7-152
Me.ace2	20091882	MW2-Y2B3A2C2D2R3	70-213
Me.bar	73668518	wm-YB1AR1D	1-144
Me.bur	91772416	MW-YB1ACDR1	8-151
Me.fla	91776283	Y4A2W3MMR1DB1Y3Z	24-172
Me.hun	88601430	Y1B1A1/C1Dc2	14-159
Me.mar	45358491	WBADMrc1c2y	3-148
Me.maz1	21226428	MW1-Y1B1A1C1D1R1	70-213
Me.maz2	21227424	w2m-Y2B2A2R2C2D2	7-152
Mo.the	83589652	DRCY	10-160
Na.pha	76801697	Y2C1D	14-160
Ni.eur	30249812	A2W3MM-RDB	21-167
Oc.ihe	23099036	BW1CD	9-159
Pe.car	77918805	A2W4R1BD---Y1X2X3	4-155
Pe.lut*	78187254	D	36-182
Ps.aer	15595372	MY1A1W1MR1DB1	24-169
Ps.syr	28868129	MY1-A1MW1R1DB1	10-158
Py.aby	14521749	wmRYBAC2C1DM	5-153
Py.hor	14590399	wm-RYBAC1C2DM	4-152
Ra.eut	73539433	M-Y1W3MM-----A2W4R4DB4Y2Z	25-173
Ra.met	94312624	M-Y3W2MM-----A2W3R2DB1Y4Z	27-175
Ra.sol	17549623	Y5A2W2MR1DBY4Z2	52-198
Rh.etl	86356295	M-Y1A1W1R1B1Y2D	10-155
Rh.fer1	89899381	Y1A1W2R1D1B1	60-202
Rh.fer2	89899713	MY2-A2MW3R2D2B2	10-159
Rh.rub	83592743	Y2A2W1MW2MMR2B2Dm	8-151
Rh.sph	77462997	Y4M-MD-Y3A2W3R3Y2	16-158
Sa.deg	90022745	mm--Y5A3W4MR3DB3	36-187
Sh.den	91794648	MY2-A2MW3R2DB2	13-161
Sh.one1	24373685	M--Y1A1W1MR1D1B1	28-175
Sh.one2	24373873	Y2--MW2R2D2B2	9-162
Si.mel	15964397	M-Y1A1W1R1B1Y2D	10-155
Silic1	99078189	y1r1w1ay2-MB1D1	11-153
Silic2	99078222	MR2D2W2	9-172
Sy.aci1	85859034	m--w3---MR2D1-Y1-D2B2A2Y2X1	4-153
Sy.aci2	85859038	m--w3---MR2D1-Y1-D2B2A2Y2X1	12-155
Sy.the	51892677	W2W3ACDYBRX1	7-158
Tb.den	74317634	Y2A1W1M-MMMRDBY1Z	24-171
Th.kod	57640574	wmRYBAAC1C2MD	3-151
Th.mar	15643665	CD	3-146
Th.ten	20807861	B2A2W4CD	6-155

Table A.5 (continued)

ID	GI	Gene Neighborhood	Range
Tm.cru	78485954	DA2	14-160
Tm.den	78777175	Y1W1M-A1R1-DB1-Z1M-M----mM	6-151
Vi.cho	15601839	Y1A1W2W1MR1DM	50-201
Vi.vul	27367544	Y2A2W4W3MR3DB3M	31-182
Xa.axo	21242633	W5-Y2A1M-MM-MMMMMMMR2DB2	23-166
Xa.cam	21231315	W6-Y2A1M--MMMM-MM-M-W1-MR2DB2	23-166
Xa.ory	77760740	W4-Y5A4-MM-M-MMMM-W5----R2DB3	23-166
Zy.mob	56550976	M-A1RB1DYW	11-156

Table A.6 CheZ data. ID, GI, Gene Neighborhood, and Range are explained in Table A.2. A neighbor-joining tree was built from the CheZ alignment in MEGA with pairwise deletions and the Poisson correction distance.

ID	GI	Gene Neighborhood	Range
Bo.bro	33601531	Y1AWMRBY2Z	14-210
Bo.par	33596132	Y1AWMRBY2Z	14-210
Bo.per	33592178	AWMBYZ	14-210
Br.jap	27382641	Z	96-267
Bu.cen	107024387	Y2A1W1MR1DB1Y1Z	29-243
Bu.mal	53724314	Y4A2W4MR2DB1-Y3Z	28-242
Bu.pse	53720909	Y2A1W2MR1DB1Y1Z	28-242
Bu.tha	83718679	Y4A3W4MR3D2B4Y5Z	28-242
Bu.xen	91785651	Y4AWMR4DB4Y3Z	23-236
Burkh	78064842	Y3A2W3MR2DB3Y4Z	29-243
Ca.cre	16124844	r2b2y4w2a2m-Y3ZM	4-153
Ca.jej	15792049	Z	32-230
Ch.sal	92114139	AWMRBMYZ	18-229
Ch.vio	34498904	d2b5r3m--w3a3y4--V2V3Y5ZA4	10-207
Co.psy	71278344	Y2ZAB---W2W1	39-243
De.aro	71906376	MY1A1W1MMR1D1B1-V1V2Y2ZA2	89-287
De.des	78357191	Z	62-241
De.vul	46579869	Z	52-238
Er.car	50120629	m-----AWMRBYZ	16-214
Es.col	15802293	AWMMRBYZ	16-214
Ha.che	83647837	Y6ZA5B5---W9W8	52-263
He.aci	109946961	Z	43-241
He.hep	32266158	Z	21-217
He.pyl	15611226	Z	44-242
Id.loi	56460223	YZAB-W2W1	42-249
La.int	94987496	Z	67-259
Ma.mag1	83312601	Z1	51-213
Ma.mag2	83312935	Z2	74-243
Ma.mag3	83313209	Z3	12-190
Me.fla	91776280	Y4A2W3MMR1DB1Y3Z	24-221
Ni.eur	30249876	Y3Z	10-206
Ni.ham	92118659	Z	89-260
Ni.win	75676755	Z	88-259
Ph.lum	37525787	AWMMRBYZ	17-216
Ph.pro	54308134	Y3ZA2B2-W1W2	36-238
Ps.aer	15596654	Y4ZA3B3---W4W5	49-262
Ps.atl	109899333	Y1ZAB-W2W1	42-247
Ps.ent	104782799	Y1ZA2B3---W4W3	48-262
Ps.flu	70729059	Y1ZA2B2---W3W4	48-262
Ps.hal	77359760	Y1ZAB---W1W2	43-253
Ps.put	26991029	Y1ZA2B3---W4W3	48-262
Ps.syr	28869185	Y2ZA3B3---W4W5	14-228
Ra.eut	73539436	M-Y1W3MM-----A2W4R4DB4Y2Z	12-210
Ra.met	94312627	M-Y3W2MM-----A2W3R2DB1Y4Z	14-211

Table A.6 (continued)

ID	GI	Gene Neighborhood	Range
Ra.sol1	17545461	Y3Z1	21-217
Ra.sol2	17549620	Y5A2W2MR1DBY4Z2	24-220
Rh.fer	89902465	Y6Z	17-215
Rh.pal	39934219	Z	92-263
Rh.rub1	83591551	Z1	9-253
Rh.rub2	83593070	Z2	85-254
Rh.rub3	83594168	Y5Z3	82-259
Rh.rub4	83594542	Z4	88-256
Rh.rub5	83595061	Z5	1-121
Sa.deg	90021807	Y4ZA2B1---W3W2	50-266
Sa.ent	16760864	AWMRBYZ	16-214
Sh.boy	82543639	RBYZ	16-214
Sh.den	91792703	Y1ZA1B1--W1W2	43-245
Sh.fle	30063333	AWMMRBYZ	16-214
Sh.one	24374720	Y4ZA2B3--W4W3	43-245
Sh.son	74311769	AWMMRBYZ	16-214
So.glo	85060042	WBZ	31-226
Tb.den	74317631	Y2A1W1M-MMMRDBY1Z	32-230
Tm.cru	78485094	Y1ZA1W2--y2w3-RBM	38-217
Tm.den1	78777178	Y1W1M-A1R1-DB1-Z1M-M----mM	2-198
Tm.den2	78778059	Z2	47-243
Vi.cho	15642064	Y4ZA3B2-W5W4	36-239
Vi.fis	59712439	YZAB--W2W1	36-238
Vi.par	28899004	YZAB-W2W1	36-246
Vi.vul	27365299	Y1ZA1B1-W1W2	36-246
Wo.suc	34558406	Z	45-244
Xa.axo	21242675	Y3ZA4	47-208
Xa.cam	21231352	Y3ZA2	47-208
Xa.ory	58582246	Y4ZA3	47-208
Ye.pes	16121943	MRBYZ	16-214
Ye.pse	51596719	m-----AW--MMRBYZ	16-214

Table A.7 CheC data. ID, GI, Gene Neighborhood, and Range are explained in Table A.2. A minimum evolution tree was built from the CheC alignment in MEGA with pairwise deletions and the Poisson correction distance.

ID	GI	Gene neighborhood	Range
An.deh	86160730	MW11Y8A7C	14-202
Ar.ful	11498644	MW-YBACDR--M	18-200
Ba.cla	56964011	BCD	8-197
Ba.hal	15614997	BCD	15-202
Ba.lic	52080248	BAWCD	15-204
Ba.sub	16078708	BAWCD	15-204
Bu.cen	107027510	Y4C	14-200
Bu.mal1	53715876	Y1C1	14-199
Bu.mal2	53717485	Y2C2	14-195
Bu.pse1	53721841	Y3C1	17-198
Bu.pse2	53723360	Y4C2	10-195
Bu.tha1	83717421	Y2C1	14-199
Bu.tha2	83718338	Y1C2	14-195
Bu.xen	91780058	Y2C	14-199
Burkh	78060866	Y1C	17-203
Ca.hyd	78043449	DR2CY1----M-MY3-R1MW1B2Y2A2	13-197
Cl.ace	15895487	W3DBR2A2CY3W2	14-196
Cl.tet	28211382	W2DBRACYW1	16-197
Co.psy1	71278577	Y1C1	14-194
Co.psy2	71281424	C2	188-370
De.haf	89895716	CD	11-195
De.psy	51245405	Y1C	10-187
Ge.kau	56419779	BAW1CD	15-204
Ge.met	78224401	R7Y7CD3	14-199
Ge.sul	39998290	R5MW10Y7A4CD3	14-199
Ha.che1	83643563	Y4C1	10-193
Ha.che2	83644505	C2	151-333
Ha.che3	83646099	C3	10-172
Ha.mar1	55377404	C1	13-195
Ha.mar2	55378056	C2	13-197
Ha.mar3	55378887	YC3D	11-188, 205-387
Ha.mar4	55379264	m---C4	13-196
Halob1	15790087	W2YBAC2C1DR	11-176, 189-367
Halob2	15790088	W2YBAC2C1DR	13-194
Halob3	15790567	C3	13-194
Id.loi	56459621	C	145-330
Le.int	24213951	W1A1/CB1Y1	888-1069
Ma.mag	83313252	Y10C--M	13-195
Me.ace1	20088911	w1m--Y1B1A1R1C1D1	15-199
Me.ace2	20091883	MW2-Y2B3A2C2D2R3	18-199
Me.bur	91772415	MW-YB1ACDR1	15-196
Me.hun1	88601429	Y1B1A1/C1Dc2	813-999
Me.hun2	88601431	C2da1/c1b1y1	349-534
Me.hun3	88602440	C3	23-209

Table A.7 (continued)

ID	GI	Gene neighborhood	Range
Me.hun4	88603918	C4--C5	11-194
Me.hun5	88603921	C4--C5	12-193
Me.mar1	45358494	YC2C1Rmdabw	10-202
Me.mar2	45358495	YC2C1Rmdabw	39-227
Me.maz1	21226429	MW1-Y1B1A1C1D1R1	18-199
Me.maz2	21227425	w2m-Y2B2A2R2C2D2	15-203
Mo.the	83589654	DRCY	22-207
My.xan	108764006	M-W13Y3A8C	14-197
Na.pha1	76801696	Y2C1D	11-185, 199-381
Na.pha2	76802199	C2-m-----m	11-188, 203-384
Oc.ihe	23099035	BW1CD	14-205
Ph.pro	54308932	x1-C	124-300
Ps.atl	109900135	C	149-332
Ps.flu	70729827	C	143-323
Ps.hal	77360153	C	145-327
Ps.syr	28868256	C	148-328
Py.aby1	14521750	wmRYBAC2C1DM	11-196
Py.aby2	33356802	wmRYBAC2C1DM	25-208
Py.hor1	14590397	wm-RYBAC1C2DM	23-206
Py.hor2	14590398	wm-RYBAC1C2DM	11-196
Rh.fer	89900674	Y5C	11-194
Sa.deg	90020042	C-M	15-197
Sh.den	91791897	C	143-325
Sh.one1	24372163	C1	143-325
Sh.one2	24374089	Y3C2	13-184
Sy.the	51892676	W2W3ACDYBRX1	11-193
Th.kod1	57640571	wmRYBAAC1C2MD	18-201
Th.kod2	57640572	wmRYBAAC1C2MD	11-196
Th.mar	15643666	CD	11-200
Th.ten	20807862	B2A2W4CD	17-200
Tm.den	78777233	Y2C	13-194
Vi.cho1	15600959	C1	144-326
Vi.cho2	15641096	Y2C2	13-188
Vi.fis1	59713791	C1	144-326
Vi.fis2	59713978	C2	144-326
Vi.par1	28898256	C1---M	143-319
Vi.par2	28901017	C2	144-326
Vi.vul1	27365859	C1---M	141-317
Vi.vul2	27367134	C2	144-326

Table A.8 CheX data. ID, GI, Gene Neighborhood, and Range are explained in Table A.2. A minimum evolution tree was built from the CheX alignment in MEGA with pairwise deletions and the Poisson correction distance.

ID	GI	Gene Neighborhood	Range
Ac.bac1	94968795	MW2R2A2B2Y2X1	3-150
Ac.bac2	94969948	Y3X2	30-182
Ba.hal	15613651	X	3-149
Bd.bac1	42523309	Y2X1	15-164
Bd.bac2	42524490	X2	158-307
Bo.bur	15595016	A2W3/R3XY3	3-157
Bo.gar	51598926	A2W3/R3XY3	3-157
Cl.ace1	15893874	X1X2	127-280
Cl.ace2	15893875	X1X2	3-150
Co.psy	71281803	X	3-152
De.aro	71908715	m----X	28-191
De.des1	78355328	X1	6-154
De.des2	78356243	R1-X2	7-155
De.haf	89893921	X	4-149
De.psy1	51246501	Y2X2X1---r2b3w2a	12-181, 282-452
De.psy2	51246502	Y2X2X1---r2b3w2a	1-148
De.vul	46578718	X	6-154
Ge.met1	78221408	X1	15-166
Ge.met2	78222816	Y1X2	135-283
Ge.met3	78224309	Y6X4X3	12-181
Ge.met4	78224310	Y6X4X3	16-164
Ge.sul1	39995243	X1	15-166
Ge.sul2	39995512	MM-Y1X2X3	16-164
Ge.sul3	39995513	MM-Y1X2X3	12-181
Ge.sul4	39996718	Y4X4	135-283
Ge.sul5	39997157	X5	1-147
Id.loi	56461032	X	3-152
Le.int	24215169	X	5-153
Pe.car1	77918182	X1	1-148
Pe.car2	77918810	A2W4R1BD---Y1X2X3	1-149
Pe.car3	77918811	A2W4R1BD---Y1X2X3	12-181
Ph.pro1	54308930	c-X1	10-173
Ph.pro2	54310407	X2	3-152
Polar	91789132	Y3X	13-155
Ps.atl	109896579	X	3-152
Ps.cry1	93005217	Y1X1X2	3-147
Ps.cry2	93005218	Y1X1X2	4-148
Ps.hal1	77359455	X1	3-152
Ps.hal2	77360876	X2	10-173
Sh.den1	91791887	X1	3-152
Sh.den2	91793175	X2	10-173
Sh.one1	24373794	mX1	10-173
Sh.one2	24375403	X2	3-152
Sy.aci1	85859042	m--w3---MR2D1-Y1-D2B2A2Y2X1	19-167

Table A.8 (continued)

ID	GI	Gene Neighborhood	Range
Sy.aci2	85860805	X2	3-151
Sy.the1	51892681	W2W3ACDYBRX1	3-150
Sy.the2	51894144	X2	3-151
Sy.the3	51894340	X3	3-150
Th.mar	15644366	X	3-151
Tm.den	78777801	X	315-464
Tr.den	42527001	AW1/R2XY	3-150
Tr.pal	15639356	AW1/R1XY	3-150
Vi.cho1	15640404	X1	3-152
Vi.cho2	15641210	X2	14-177
Vi.fis1	59710912	X1	3-152
Vi.fis2	59712133	X2	14-177
Vi.par1	28898043	X1	10-173
Vi.par2	28899504	X2	3-152
Vi.vul1	27364791	X1	1-133
Vi.vul2	27365703	X2	10-173
Wo.suc	34556651	X	321-470

Table A.9 CheV data. ID, GI, Gene Neighborhood, and Range are explained in Table A.2. A minimum evolution tree was built from the CheV alignment in MEGA with pairwise deletions and the Poisson correction distance.

ID	GI	Gene Neighborhood	Range
Ba.ant	49184586	V	14-142
Ba.cer	30019800	V	14-301
Ba.cla	56963907	V	13-287
Ba.hal1	15614275	V1	11-296
Ba.hal2	15614825	V2	15-302
Ba.lic	52080002	V	15-302
Ba.sub	16078465	M-----V	15-301
Ba.thu	49477317	V	14-301
Bu.xen	91782807	V	18-316
Ca.jej	15791655	VAW	12-309
Ch.vio1	34498487	V1	18-311
Ch.vio2	34498901	d2b5r3m--w3a3y4--V2V3Y5ZA4	18-311
Ch.vio3	34498902	d2b5r3m--w3a3y4--V2V3Y5ZA4	22-311
Cl.ace	15894516	V	12-298
Co.psy	71278949	V-R	18-305
De.aro1	71906373	MY1A1W1MMR1D1B1-V1V2Y2ZA2	18-312
De.aro2	71906374	MY1A1W1MMR1D1B1-V1V2Y2ZA2	23-309
De.des1	78355438	V1	13-312
De.des2	78356538	V2	13-312
De.des3	78358108	V3-----m	14-308
De.des4	78358140	V4	13-308
De.vul1	46578473	V1	13-313
De.vul2	46579066	V2	13-311
De.vul3	46579405	V3	13-314
Er.car	50120503	V	18-316
Ge.met	78223935	V	13-311
Ge.sul	39995985	V	13-311
Ha.che1	83647407	V1-R5	18-306
Ha.che2	83647449	V2	18-304
He.aci1	109947053	V1AW	13-308
He.aci2	109947387	V2	15-312
He.aci3	109948162	V3	18-318
He.hep1	32266170	V1AW	14-314
He.hep2	32266318	Y-----V2	19-319
He.hep3	32267301	V3	18-317
He.pyl1	15611088	V1	18-318
He.pyl2	15611626	V2	15-312
He.pyl3	15612053	V3AW	13-308
Id.loi	56460257	VR	18-305
La.int1	94987171	V1	13-311
La.int2	94987559	V2	13-308
Li.inn	16799772	R-----V-YA	14-301
Li.mon	16802731	R-----V-YA	14-301
Pe.car	77919641	V	16-315

Table A.9 (continued)

ID	GI	Gene Neighborhood	Range
Ph.pro1	54302441	V1Y1	18-129
Ph.pro2	54308092	V2R2	18-303
Ph.pro3	54309096	V3	8-292
Ps.aer	15598545	VR3	18-304
Ps.atl	109899407	VR	18-303
Ps.ent1	104780172	V1---m	9-291
Ps.ent2	104782743	V2	18-304
Ps.ent3	104782849	V3R4	18-305
Ps.flu1	70731554	V1	18-304
Ps.flu2	70731828	V2R3	18-304
Ps.flu3	70732365	V3---m	25-308
Ps.hal	77359716	VR	18-308
Ps.put1	26987538	V1	9-291
Ps.put2	26988852	V2	18-304
Ps.put3	26991082	V3R3	18-305
Ps.syr1	28868531	V1	24-306
Ps.syr2	28869131	V2R3	18-304
Ps.syr3	28870685	V3	18-304
Ra.eut	73538810	V	18-316
Ra.met	94313855	V	18-316
Ra.sol	17548819	V	16-312
Rh.fer	89899178	V	18-315
Sa.deg	90021860	VR1	18-304
Sa.ent	16761240	V	18-316
Sh.den1	91792659	V1R1	18-304
Sh.den2	91793498	V2	18-308
Sh.den3	91793879	V3	7-291
Sh.one1	24373553	V1	18-308
Sh.one2	24374641	V2	7-291
Sh.one3	24374763	V3R3	18-304
Tb.den	74317616	V	18-313
Tm.cru1	78485615	V1	18-319
Tm.cru2	78485817	V2	18-312
Tm.den1	78777455	V1	17-311
Tm.den2	78777728	V2A2W2	12-308
Vi.cho1	15601707	V1	10-295
Vi.cho2	15641610	V2	18-325
Vi.cho3	15642008	V3	18-309
Vi.cho4	15642201	V4R3	18-306
Vi.fis1	59712486	V1R	18-304
Vi.fis2	59713881	V2	7-291
Vi.fis3	59713985	V3	18-309
Vi.par1	28897547	V1R	18-306
Vi.par2	28898811	V2	18-309
Vi.par3	28900286	V3	10-294
Vi.par4	28900601	V4	15-322

Table A.9 (continued)

ID	GI	Gene Neighborhood	Range
Vi.vul1	27363710	V1R1	18-306
Vi.vul2	27366264	V2	18-309
Vi.vul3	27366527	V3---m	15-322
Vi.vul4	27367961	MM---V4	1-281
Wo.suc1	34556697	V1	17-317
Wo.suc2	34558094	V2	35-330
Wo.suc3	34558164	V3	18-299
Wo.suc4	34558368	V4AW3	15-310
Xa.axo	21242731	V	19-307
Xa.cam	21231401	V	19-307
Xa.ory	58582191	V	19-307

Table A.10 CheW data. ID, GI, Gene Neighborhood, and Range are explained in Table A.2. A minimum evolution tree was built from the CheW alignment in MEGA with pairwise deletions and the Poisson correction distance.

ID	GI	Gene Neighborhood	Range
Ac.bac1	94968553	A1W1MY1B1R1	10-143
Ac.bac2	94968800	MW2R2A2B2Y2X1	6-139
Acine	50084003	Y1Y2W-A	28-161
Ag.tum1	15889359	W1	42-175
Ag.tum2	15889873	W2M	54-187
An.deh1	86157030	r2b2a2mw3w2Y3--Y2-D1A1W1R1B1	12-149
An.deh2	86157038	b1r1w1a1d1-y2--y3W2W3MA2B2R2	15-147
An.deh3	86157039	b1r1w1a1d1-y2--y3W2W3MA2B2R2	1-120
An.deh4	86157621	R3W4MA3B3	65-195
An.deh5	86157801	b5r5w6ma5y5--W5MB4d2r4a4	25-163
An.deh6	86157807	A4R4D2b4mw5--Y5A5MW6R5B5	11-148
An.deh7	86158608	MW7	13-150
An.deh8	86159155	A6-W9W8-R6B6Y7	15-158
An.deh9	86159156	A6-W9W8-R6B6Y7	5-140
An.deh10	86160371	W10R9	43-176
An.deh11	86160733	MW11Y8A7C	2-127
An.var1	75906285	Y1Y2W1MA1	21-157
An.var2	75906730	Y4Y3W2MA2	30-169
An.var3	75909981	Y6Y7W3MA3	7-148
Ar.ful	11498649	MW-YBACDR--M	17-151
Azoar	56476383	Y1Y2WMA	28-160
Ba.cla	56964346	AW	6-129
Ba.hal	15615531	AWY2	13-146
Ba.lic	52080247	BAWCD	10-143
Ba.sub	16078707	BAWCD	10-143
Bd.bac1	42522174	A1W1D1/B1	122-253
Bd.bac2	42524237	MR1D2W2	9-144, 172-307, 341-483
Bd.bac3	42524824	b2r2a2-W3	12-145
Bo.bro	33601526	Y1AWMRBY2Z	18-153
Bo.bur1	15594657	W1	15-160
Bo.bur2	15594910	W2-A1B2-Y2	13-159
Bo.bur3	15595015	A2W3/R3XY3	28-161
Bo.gar1	51598572	W1	15-160
Bo.gar2	51598816	W2-A1B2-Y2	13-159
Bo.gar3	51598925	A2W3/R3XY3	28-161
Bo.par	33596127	Y1AWMRBY2Z	24-159
Bo.per	33592174	AWMBYZ	18-153
Br.jap1	27375503	A1W1Y1R1	14-147
Br.jap2	27377304	A2W2Y2B1R2	14-147
Br.jap3	27377455	Y3A3W3MW4MR3B2	16-153
Br.jap4	27377457	Y3A3W3MW4MR3B2	16-153
Bu.cen1	107024393	Y2A1W1MR1DB1Y1Z	31-164
Bu.cen2	107026916	MW3R2-W2A2B3	85-225
Bu.cen3	107026919	MW3R2-W2A2B3	13-150

Table A.10 (continued)

ID	GI	Gene Neighborhood	Range
Bu.mal1	53716489	MW2R1W1A1B2	54-193
Bu.mal2	53716497	MW2R1W1A1B2	2-146
Bu.mal3	53723386	MW3	12-149
Bu.mal4	53724321	Y4A2W4MR2DB1-Y3Z	29-162
Bu.pse1	53719345	MW1	24-161
Bu.pse2	53720915	Y2A1W2MR1DB1Y1Z	29-162
Bu.pse3	53722892	MW4R2W3A2B2	97-236
Bu.pse4	53722894	MW4R2W3A2B2	10-154
Bu.tha1	83717195	MW1R2W3A2B3	2-146
Bu.tha2	83717542	Y3A1W2MR1D1B2B1	10-145
Bu.tha3	83718252	MW1R2W3A2B3	92-231
Bu.tha4	83719095	Y4A3W4MR3D2B4Y5Z	29-162
Bu.tha5	83720231	MW5	12-149
Bu.xen	91785657	Y4AWMR4DB4Y3Z	32-165
Burkh1	78063152	MW2R1W1A1B2	87-227
Burkh2	78063154	MW2R1W1A1B2	13-150
Burkh3	78064836	Y3A2W3MR2DB3Y4Z	25-158
Ca.cre1	16124689	M-M-Y1A1W1R1B1Y2D	14-147
Ca.cre2	16124849	mzy3-MA2W2Y4B2R2	5-139
Ca.cre3	16125017	W3	10-142
Ca.cre4	16127255	W4	14-148
Ca.hyd1	78044189	DR2CY1----M-MY3-R1MW1B2Y2A2	7-140, 174-314, 342-478
Ca.hyd2	78044727	MW2A1B1	12-145
Ca.jej	15791653	VAW	31-169
Ch.sal	92114145	AWMRBMYZ	18-151
Ch.vio1	34496467	Y1-A1MW1MD1B1	24-161
Ch.vio2	34497963	MY2-A2MW2R2B3	18-155
Ch.vio3	34498896	a4zy5v3v2--Y4A3W3--MR3B5D2	18-151
Cl.ace1	15893415	Y1A1W1MR1Y2	18-150
Cl.ace2	15895485	W3DBR2A2CY3W2	2-130
Cl.ace3	15895492	W3DBR2A2CY3W2	6-139
Cl.tet1	28211380	W2DBRACYW1	2-131
Cl.tet2	28211387	W2DBRACYW1	7-142
Co.psy1	71278201	Y2ZAB---W2W1	20-153
Co.psy2	71278336	Y2ZAB---W2W1	95-226
De.aro1	71906366	MY1A1W1MMR1D1B1-V1V2Y2ZA2	20-153
De.aro2	71906710	W2	716-863
De.aro3	71906783	Y3M-A3MW3D2R2D3B2	22-159
De.aro4	71909508	Y7Y6W4MA4	29-160
De.des1	78355622	W1	14-151
De.des2	78356125	MW2	6-143
De.des3	78356618	B1-R2-W3Y3A1	100-233
De.des4	78357083	W4	14-147
De.haf1	89895706	W1	4-139
De.haf2	89895739	m--AW2B	12-144

Table A.10 (continued)

ID	GI	Gene Neighborhood	Range
De.haf3	89896487	W3M	2-134
De.psy1	51246211	W1M	6-144, 171-303, 343-480
De.psy2	51246495	AW2B3R2---x1x2y2	13-146
De.rad	15808010	YWMMMA	4-135
De.vul1	46579006	MW1	6-143
De.vul2	46580313	W2	14-147
De.vul3	46580370	MW3A2	11-148
De.vul4	46580479	B3-R3-W4Y3A3	45-178
Er.car	50120624	m-----AWMRBYZ	18-151
Er.lit	85375079	AWYBR-M	5-132
Es.col	15802299	AWMMRBYZ	18-151
Ge.kau1	56419778	BAW1CD	10-143
Ge.kau2	56420881	W2	3-135
Ge.met1	78222297	A1W1MR2D1B2	11-144
Ge.met2	78223510	Y3A2-W3-W2-R3B3-Y2	10-143
Ge.met3	78223512	Y3A2-W3-W2-R3B3-Y2	100-233
Ge.met4	78223624	Y5--Y4A3--MMMW4R4D2B4	25-162
Ge.met5	78223902	W5R6W6MA4B6	6-137
Ge.met6	78223904	W5R6W6MA4B6	64-194
Ge.met7	78224456	W7A5R8-B8-R9	18-152
Ge.met8	78224508	W8	19-154
Ge.sul1	39995406	W1A1R2-B1-R1	18-152
Ge.sul2	39995790	MW2	11-148
Ge.sul3	39996244	MMW3R3D2B2	11-148
Ge.sul4	39996401	Y2M-Y3A2---M---MW4MW5-MM	11-148
Ge.sul5	39996403	Y2M-Y3A2---M---MW4MW5-MM	11-148
Ge.sul6	39997316	Y6A3-W7-W6--R4B3-Y5	9-142
Ge.sul7	39997318	Y6A3-W7-W6--R4B3-Y5	21-154
Ge.sul8	39997511	W8	19-154
Ge.sul9	39997673	MW9	1-111
Ge.sul10	39998287	R5MW10Y7A4CD3	7-131
Gl.oxy	58039986	M-Y1AWRBY2	1-127
Ha.che1	83643356	M----Y1A1W1MMW2-R1D1B1	24-161
Ha.che2	83643359	M----Y1A1W1MMW2-R1D1B1	16-157
Ha.che3	83643434	Y3Y2W4MR2A2B2W3	10-148
Ha.che4	83643439	Y3Y2W4MR2A2B2W3	36-170
Ha.che5	83646433	Y5-A3-MW5R3D2B3	16-153
Ha.che6	83646567	W7R4W6MA4B4	59-189
Ha.che7	83646569	W7R4W6MA4B4	8-139
Ha.che8	83647830	Y6ZA5B5---W9W8	17-150
Ha.che9	83647831	Y6ZA5B5---W9W8	290-420
Ha.che10	83648469	Y7Y8W10-MA6-W11	52-196
Ha.che11	83648474	Y7Y8W10-MA6-W11	10-143
Ha.mar1	55378265	W1	187-319
Ha.mar2	55378895	rabW2	10-140
Halob1	15790067	W1	7-132

Table A.10 (continued)

ID	GI	Gene Neighborhood	Range
Halob2	15790092	W2YBAC2C1DR	15-165
He.aci	109947051	V1AW	29-161
He.hep	32266172	V1AW	27-159
He.pyl	15612055	V3AW	29-161
Id.loi1	56460218	YZAB-W2W1	22-155
Id.loi2	56460219	YZAB-W2W1	140-272
Janna	89055331	db2-Y2AWR2Y1	14-147
La.int1	94987583	B-R-W1Y2AA	14-148
La.int2	94987612	W2	17-150
Le.int1	24213950	W1A1/CB1Y1	37-170
Le.int2	24214221	W2	8-141
Le.int3	24215127	Y2-A2MW3D1B3	9-150
Ma.mag1	83309424	A1W1Y1B1R1	22-155
Ma.mag2	83311061	M--W2R2W3MB2	11-144
Ma.mag3	83311063	M--W2R2W3MB2	54-184
Ma.mag4	83312106	MB4R4A2--W4	42-175
Me.ace1	20088919	d1c1r1a1b1y1--MW1	19-153
Me.ace2	20091888	MW2-Y2B3A2C2D2R3	38-172
Me.bar	73668525	dr1ab1y-MW	18-152
Me.bur	91772410	MW-YB1ACDR1	15-149
Me.cap1	53805036	W2R2W1AB2	75-209
Me.cap2	53805038	W2R2W1AB2	40-176
Me.fla1	91775597	Y1Y2W1MA1	32-165
Me.fla2	91776113	W2	709-847
Me.fla3	91776287	Y4A2W3MMR1DB1Y3Z	26-159
Me.hun1	88601325	MM----MW2---W1	49-193
Me.hun2	88601329	MM----MW2---W1	20-164
Me.hun3	88601799	m--W3MA2	13-143, 203-332
Me.hun4	88602189	W4	38-181
Me.hun5	88602251	W5R2	622-772
Me.hun6	88602282	W7R3W6MA3B4	38-168
Me.hun7	88602284	W7R3W6MA3B4	21-150
Me.hun8	88602699	MW8	38-182
Me.hun9	88602914	MW9-----m	1-121
Me.hun10	88603177	W10	22-165
Me.hun11	88603642	W11	40-183
Me.hun12	88603773	w13W12	6-149
Me.hun13	88603774	w12W13	10-159
Me.hun14	88603790	mW14	4-157
Me.lot1	13488380	MW2RW1AB	85-226
Me.lot2	13488382	MW2RW1AB	4-140
Me.mar	45358488	WBADMrc1c2y	6-142
Me.maz1	21226434	MW1-Y1B1A1C1D1R1	15-149
Me.maz2	21227432	d2c2r2a2b2y2-MW2	1-133
Mo.the	83589593	MWAB	10-143
My.xan1	108756921	A7W1M-R6B2--Y1---b3a3mw3r2w5	4-137

Table A.10 (continued)

ID	GI	Gene Neighborhood	Range
My.xan2	108757829	Y8W2R5W6A2B5M	131-267
My.xan3	108758502	W5R2W3MA3B3---y1--b2r6-mw1a7	57-185
My.xan4	108759499	Y7A1W9W4--R8B1Y5	28-161
My.xan5	108759678	W5R2W3MA3B3---y1--b2r6-mw1a7	7-138
My.xan6	108759864	Y8W2R5W6A2B5M	41-170
My.xan7	108759985	W7-MMA5-B6R4	12-138
My.xan8	108760558	W10R7MY4W8A6	59-184
My.xan9	108761164	Y7A1W9W4--R8B1Y5	43-186
My.xan10	108761716	W10R7MY4W8A6	48-182
My.xan11	108762907	W14W11MA4B7R3	2-98
My.xan12	108763064	Y6-----W12	1-127, 151-270
My.xan13	108763081	M-W13Y3A8C	2-129
My.xan14	108764028	W14W11MA4B7R3	16-150
Na.pha	76802707	W	13-141
Ni.eur1	30249009	Y1W1	29-160
Ni.eur2	30249368	W2	9-141, 172-318, 357-496
Ni.eur3	30249817	A2W3MM-RDB	23-156
Ni.ham	92118836	AWY2B2R2	14-147
Ni.mul1	82701463	W1RW2MAB	12-143
Ni.mul2	82701465	W1RW2MAB	57-187
Ni.oce1	77163662	Y2Y1W2MR1AB1W1	41-176
Ni.oce2	77163667	Y2Y1W2MR1AB1W1	35-168
Ni.win	75674719	AWY1BR-----M	14-147
Nosto1	17228423	Y2Y1W1MA1	30-169
Nosto2	17228565	Y4Y3W2MA2	16-157
Nosto3	17229655	Y6Y5W3MA3	21-157
Oc.ihe1	23099034	BW1CD	10-143
Oc.ihe2	23099997	AW2	6-139
Pe.car1	77917648	MW1A1	2-131
Pe.car2	77918161	W2m	36-169
Pe.car3	77918601	W3	21-154
Pe.car4	77918802	A2W4R1BD---Y1X2X3	24-156
Pe.car5	77919957	MW5	22-155
Ph.lum	37525781	AWMMRBYZ	18-151
Ph.pro1	54308138	Y3ZA2B2-W1W2	230-360
Ph.pro2	54308139	Y3ZA2B2-W1W2	20-153
Polar1	91787036	Y1Y2W1MA1	29-166
Polar2	91788323	W2RW3--MMMA2B	2-133
Polar3	91788325	W2RW3--MMMA2B	38-168
Ps.aer1	15595375	MY1A1W1MR1DB1	15-148
Ps.aer2	15595607	Y2Y3W2MR2A2B2W3	36-168
Ps.aer3	15595612	Y2Y3W2MR2A2B2W3	25-159
Ps.aer4	15596660	Y4ZA3B3---W4W5	150-281
Ps.aer5	15596661	Y4ZA3B3---W4W5	15-148
Ps.aer6	15598900	MW7R4W6A4B4	82-221
Ps.aer7	15598902	MW7R4W6A4B4	16-153

Table A.10 (continued)

ID	GI	Gene Neighborhood	Range
Ps.arc	71066371	Y2Y1WMRA	32-166
Ps.atl1	109899328	Y1ZAB-W2W1	20-153
Ps.atl2	109899329	Y1ZAB-W2W1	144-275
Ps.cry	93006923	Y3Y2WMRAB	32-166
Ps.ent1	104780452	MW1R2W2A1B1	18-155
Ps.ent2	104780454	MW1R2W2A1B1	79-213
Ps.ent3	104782792	Y1ZA2B3---W4W3	15-148
Ps.ent4	104782793	Y1ZA2B3---W4W3	161-292
Ps.ent5	104783968	Y3Y2W6MA3W5	13-144
Ps.ent6	104783971	Y3Y2W6MA3W5	41-175
Ps.flu1	70728512	MW1R1W2A1B1	18-155
Ps.flu2	70728514	MW1R1W2A1B1	81-220
Ps.flu3	70729065	Y1ZA2B2---W3W4	140-271
Ps.flu4	70729066	Y1ZA2B2---W3W4	16-149
Ps.flu5	70733107	Y3Y2W6MA3W5	10-142
Ps.flu6	70733110	Y3Y2W6MA3W5	36-169
Ps.hal1	77359766	Y1ZAB---W1W2	96-227
Ps.hal2	77359767	Y1ZAB---W1W2	21-154
Ps.put1	26988222	MW1R1W2A1B1	18-150
Ps.put2	26988224	MW1R1W2A1B1	73-207
Ps.put3	26991022	Y1ZA2B3---W4W3	15-148
Ps.put4	26991023	Y1ZA2B3---W4W3	151-282
Ps.put5	26991664	Y3Y2W6MA3W5	13-144
Ps.put6	26991667	Y3Y2W6MA3W5	41-174
Ps.syr1	28868131	MY1-A1MW1R1DB1	19-157
Ps.syr2	28868701	MW2R2W3A2B2	18-155
Ps.syr3	28868703	MW2R2W3A2B2	82-221
Ps.syr4	28869191	Y2ZA3B3---W4W5	149-280
Ps.syr5	28869192	Y2ZA3B3---W4W5	15-148
Ps.syr6	28869637	MW6-----m	1-131, 160-294, 325-470
Ps.syr7	28872143	Y4Y3W8MA4W7	15-147
Ps.syr8	28872146	Y4Y3W8MA4W7	36-169
Py.aby	14521757	mdc1c2abyrMW	5-141
Py.hor	14590390	mdc2c1abyr-MW	5-141
Ra.eut1	73537485	MW2R1W1A1B1	78-217
Ra.eut2	73537487	MW2R1W1A1B1	3-141
Ra.eut3	73539422	M-Y1W3MM-----A2W4R4DB4Y2Z	13-146
Ra.eut4	73539431	M-Y1W3MM-----A2W4R4DB4Y2Z	18-151
Ra.eut5	73542306	Y5Y4W5MA3	37-169
Ra.met1	94309616	Y1Y2W1MA1	37-167
Ra.met2	94312613	M-Y3W2MM-----A2W3R2DB1Y4Z	35-168
Ra.met3	94312622	M-Y3W2MM-----A2W3R2DB1Y4Z	18-151
Ra.met4	94312898	MW5R3W4A3B2	86-242
Ra.met5	94312900	MW5R3W4A3B2	3-142
Ra.sol1	17545389	Y1Y2W1MA1	36-168
Ra.sol2	17549626	Y5A2W2MR1DBY4Z2	16-149

Table A.10 (continued)

ID	GI	Gene Neighborhood	Range
Rh.etl1	86356291	M-Y1A1W1R1B1Y2D	14-147
Rh.etl2	86358434	MW2	14-147
Rh.etl3	86359108	Y4A2W4MMMW3R2B2	9-146
Rh.etl4	86359112	Y4A2W4MMMW3R2B2	21-158
Rh.etl5	86359435	W5	15-148
Rh.fer1	89899282	W1	698-835
Rh.fer2	89899379	Y1A1W2R1D1B1	17-150
Rh.fer3	89899711	MY2-A2MW3R2D2B2	26-163
Rh.fer4	89899731	W4M	28-161
Rh.fer5	89899795	W5M	28-161
Rh.fer6	89900167	Y3Y4W6MA3	29-166
Rh.fer7	89900269	W7---m	40-173
Rh.pal1	39933217	Y1A1W2W1MR1B1	6-143
Rh.pal2	39933218	Y1A1W2W1MR1B1	18-155
Rh.pal3	39934698	A2W3Y3B2R2	24-157
Rh.pal4	39934747	MA3W4R3	24-157
Rh.rub1	83592736	Y2A2W1MW2MMR2B2Dm	11-148
Rh.rub2	83592738	Y2A2W1MW2MMR2B2Dm	12-149
Rh.rub3	83593657	MW3B4-R4Y3A3--Y4--M	34-168, 206-349, 385-520
Rh.rub4	83593897	W4	29-162
Rh.sph1	77462125	Y1A1W1W2R1B1	14-151
Rh.sph2	77462126	Y1A1W1W2R1B1	23-161
Rh.sph3	77462993	Y4M-MD-Y3A2W3R3Y2	12-145
Rh.sph4	77463617	A3R4B3W4-MY5A4	25-156
Sa.deg1	90020170	Y2Y1W1MA1	47-182
Sa.deg2	90021800	Y4ZA2B1---W3W2	17-150
Sa.deg3	90021801	Y4ZA2B1---W3W2	242-372
Sa.deg4	90022748	mm--Y5A3W4MR3DB3	33-170
Sa.deg5	90023267	Y7Y6W6MR4A4B4W5	16-151
Sa.deg6	90023272	Y7Y6W6MR4A4B4W5	35-169
Sa.ent	16760869	AWMRBYZ	18-151
Sa.rub	83814312	WR2Y-AB1-M-M-M	31-164
Sh.den1	91792708	Y1ZA1B1--W1W2	196-327
Sh.den2	91792709	Y1ZA1B1--W1W2	20-153
Sh.den3	91794646	MY2-A2MW3R2DB2	22-159
Sh.fle	30063339	AWMMRB-Z	18-151
Sh.one1	24373682	M--Y1A1W1MR1D1B1	20-153
Sh.one2	24373871	Y2--MW2R2D2B2	18-155
Sh.one3	24374714	Y4ZA2B3--W4W3	20-153
Sh.one4	24374715	Y4ZA2B3--W4W3	181-312
Sh.son	74311763	AWMMRBYZ	18-151
Si.mel1	15964393	M-Y1A1W1R1B1Y2D	14-147
Si.mel2	15965904	MW2	15-148
Si.mel3	15966756	W3--M	16-149
Si.mel4	16263299	R2W4MA2B2	65-205
Silic1	99078183	d1b1m-Y2AW1R1Y1	21-154

Table A.10 (continued)

ID	GI	Gene Neighborhood	Range
Silic2	99078223	MR2D2W2	47-183, 209-347, 380-522
So.glo	85060044	WBZ	18-147
Sp.ala	103487224	AWYBR	5-130
Sy.aci1	85858533	W1R1W2MA1B1	6-137
Sy.aci2	85858535	W1R1W2MA1B1	52-182
Sy.aci3	85859028	x1y2a2b2d2-y1-d1r2m---W3--M	22-159
Sy.elo1	56750539	W1MA1	16-153
Sy.elo2	56750689	Y3Y2W3MA2W2	22-162
Sy.elo3	56750692	Y3Y2W3MA2W2	22-158
Sy.the1	51892494	W1	8-141
Sy.the2	51892673	W2W3ACDYBRX1	10-147
Sy.the3	51892674	W2W3ACDYBRX1	13-146
Syncy1	16329619	Y1Y2W1M	36-180
Syncy2	16329792	Y4Y3W2MA1	8-138
Syncy3	16331985	Y6Y5W4MMA3W3	23-150
Syncy4	16331989	Y6Y5W4MMA3W3	11-170
Synco1	86606800	Y3Y2W1---MA1	19-170
Synco2	86607344	Y4Y5Y6W2MA2	9-148
Tb.den1	74317641	Y2A1W1M-MMMRDBY1Z	21-154
Tb.den2	74318568	Y4Y3W2MA2	29-160
Th.elo1	22297890	Y1Y2W1MA1	30-169
Th.elo2	22298110	Y4Y3W3MA2W2	17-149
Th.elo3	22298113	Y4Y3W3MA2W2	23-159
Th.elo4	22298567	Y6Y5W4MA3	15-154
Th.kod	57640564	dmc2c1aabyrMW	5-141
Th.mar1	15643464	AW1Y	12-142
Th.mar2	15643481	W2	6-139
Th.ten1	20807187	W1-M	4-136
Th.ten2	20807513	MW2B1-R1Y1A1	16-149, 188-328, 359-495
Th.ten3	20807596	W3	4-137
Th.ten4	20807863	B2A2W4CD	3-137
Tm.cru1	78485073	W1	7-165
Tm.cru2	78485096	Y1ZA1W2--y2w3-RBM	32-165
Tm.cru3	78485100	mbr-W3Y2--w2a1zy1	21-160
Tm.den1	78777169	Y1W1M-A1R1-DB1-Z1M-M----mM	19-152
Tm.den2	78777726	V2A2W2	27-159
Tr.den1	42527000	AW1/R2XY	31-166
Tr.den2	42527097	W2	20-157
Tr.pal1	15639355	AW1/R1XY	40-175
Tr.pal2	15639430	W2	22-159
Vi.cho1	15601842	Y1A1W2W1MR1DM	5-139
Vi.cho2	15601843	Y1A1W2W1MR1DM	29-164
Vi.cho3	15641413	mm-MW3B1-R2Y3A2-M	19-152, 192-334, 370-503
Vi.cho4	15642059	Y4ZA3B2-W5W4	30-163
Vi.cho5	15642060	Y4ZA3B2-W5W4	194-324
Vi.fis1	59712433	YZAB--W2W1	20-153

Table A.10 (continued)

ID	GI	Gene Neighborhood	Range
Vi.fis2	59712434	YZAB--W2W1	126-256
Vi.par1	28898999	YZAB-W2W1	20-153
Vi.par2	28899000	YZAB-W2W1	227-357
Vi.vul1	27365303	Y1ZA1B1-W1W2	210-340
Vi.vul2	27365304	Y1ZA1B1-W1W2	20-153
Vi.vul3	27367547	Y2A2W4W3MR3DB3M	5-138
Vi.vul4	27367548	Y2A2W4W3MR3DB3M	24-159
Wo.suc1	34556548	W1M	13-146
Wo.suc2	34557460	Y2W2	12-145
Wo.suc3	34558366	V4AW3	26-158
Xa.axo1	21243180	MW1	18-155
Xa.axo2	21243594	A2MW2-R3B3	19-156
Xa.axo3	21243823	Y6Y5W4MA3B4W3	10-146
Xa.axo4	21243827	Y6Y5W4MA3B4W3	34-167
Xa.axo5	77748621	W5-Y2A1M-MM-MMMMMMMR2DB2	289-423
Xa.cam1	21231319	W6-Y2A1M--MMMM-MM-M-W1-MR2DB2	13-146
Xa.cam2	21231751	MW2	18-155
Xa.cam3	21232133	A3MW3-R3B4	19-156
Xa.cam4	21232350	Y6Y5W5MA4B5W4	10-146
Xa.cam5	21232354	Y6Y5W5MA4B5W4	34-167
Xa.cam6	77747856	W6-Y2A1M--MMMM-MM-M-W1-MR2DB2	276-410
Xa.ory1	58581091	A1MW1-R1B1	19-156
Xa.ory2	58581371	Y6Y2W2MA2A2B2W3	34-167
Xa.ory3	58581376	Y6Y2W2MA2A2B2W3	10-146
Xa.ory4	58582456	W4-Y5A4-MM-M-MMMM-W5----R2DB3	205-339
Xa.ory5	58582473	W4-Y5A4-MM-M-MMMM-W5----R2DB3	1-121
Xy.fas1	15838544	Y2W2MABW1	10-146
Xy.fas2	15838548	Y2W2MABW1	33-166
Ye.pes	16121931	AW	18-151
Ye.pse	51596727	m-----AW--MMRBYZ	18-151
Zy.mob	56550974	M-A1RB1DYW	25-158

Table A.11 CheY data. ID, GI, Gene Neighborhood, and Range are explained in Table A.2. Class corresponds to the flagellar (Fla) or Tfp CheY proteins identified in phylogenetic analyses. A minimum evolution tree was built from the CheY alignment in MEGA with pairwise deletions and the pairwise distance.

ID	GI	Gene Neighborhood	Class	Range
Ac.bac1	94968555	A1W1MY1B1R1	Fla	3-117
Ac.bac2	94968796	MW2R2A2B2Y2X1	Fla	18-131
Ac.bac3	94969947	Y3X2	Fla	6-123
Ac.bac4	94971681	Y4	Tfp	1-112
Acine1	50084001	Y1Y2W-A	Tfp	10-124
Acine2	50084002	Y1Y2W-A	Tfp	4-118
Ag.tum1	15887864	M-Y1ARBY2D-----M	Fla	5-119
Ag.tum2	15887868	M-Y1ARBY2D-----M	Fla	9-124
An.deh1	86156530	Y1	Fla	10-125
An.deh2	86157034	r2b2a2mw3w2Y3--Y2-D1A1W1R1B1	Tfp	6-120
An.deh3	86157037	r2b2a2mw3w2Y3--Y2-D1A1W1R1B1	Tfp	76-189
An.deh4	86157522	Y4	Tfp	12-126
An.deh5	86157804	A4R4D2b4mw5--Y5A5MW6R5B5	Fla	5-119
An.deh6	86159015	Y6	Tfp	6-120
An.deh7	86159151	A6-W9W8-R6B6Y7	Fla	9-124
An.deh8	86160732	MW11Y8A7C	Fla	5-118
An.var1	75906283	Y1Y2W1MA1	Tfp	231-345
An.var2	75906284	Y1Y2W1MA1	Tfp	9-123
An.var3	75906731	Y4Y3W2MA2	Tfp	4-118
An.var4	75906732	Y4Y3W2MA2	Tfp	276-398
An.var5	75909373	Y5	Fla	11-125
An.var6	75909979	Y6Y7W3MA3	Tfp	261-375
An.var7	75909980	Y6Y7W3MA3	Tfp	4-118
Ar.ful	11498647	MW-YBACDR--M	Fla	4-117
Azoar1	56476381	Y1Y2WMA	Tfp	8-122
Azoar2	56476382	Y1Y2WMA	Tfp	6-120
Ba.ant	49184556	Y-A---r2	Fla	5-119
Ba.cer	30019774	YA---r2	Fla	5-119
Ba.cla	56964021	Y	Fla	5-118
Ba.hal1	15615007	Y1	Fla	5-118
Ba.hal2	15615530	AWY2	Fla	4-117
Ba.lic1	52080236	Y1	Fla	5-118
Ba.lic2	52080416	Y2	Fla	4-117
Ba.sub1	16078696	Y1	Fla	5-118
Ba.sub2	16078857	Y2	Fla	4-117
Ba.thu	49480952	YA---r2	Fla	5-119
Bd.bac1	42521746	Y1	Fla	5-118
Bd.bac2	42523310	Y2X1	Fla	21-141
Bd.bac3	42523853	Y3	Fla	8-128
Bd.bac4	42524789	Y4	Fla	10-130
Bo.bro1	33601524	Y1AWMRBY2Z	Fla	5-119
Bo.bro2	33601530	Y1AWMRBY2Z	Fla	8-123
Bo.bur1	15594896	Y1	Fla	14-132

Table A.11 (continued)

ID	GI	Gene Neighborhood	Class	Range
Bo.bur2	15594915	W2-A1B2-Y2	Fla	5-123
Bo.bur3	15595017	A2W3/R3XY3	Fla	28-143
Bo.gar1	51598802	Y1	Fla	7-125
Bo.gar2	51598821	W2-A1B2-Y2	Fla	5-123
Bo.gar3	51598927	A2W3/R3XY3	Fla	28-143
Bo.par1	33596125	Y1AWMRBY2Z	Fla	5-119
Bo.par2	33596131	Y1AWMRBY2Z	Fla	8-123
Bo.per	33592177	AWMBYZ	Fla	8-123
Br.jap1	27375502	A1W1Y1R1	Fla	1-113
Br.jap2	27377305	A2W2Y2B1R2	Fla	4-118
Br.jap3	27377453	Y3A3W3MW4MR3B2	Fla	4-118
Br.jap4	27382590	Y4	Fla	9-124
Bu.cen1	107024388	Y2A1W1MR1DB1Y1Z	Fla	7-122
Bu.cen2	107024395	Y2A1W1MR1DB1Y1Z	Fla	5-119
Bu.cen3	107026642	M--Y3	Fla	4-119
Bu.cen4	107027509	Y4C	Fla	5-118
Bu.mal1	53715875	Y1C1	Fla	67-180
Bu.mal2	53717484	Y2C2	Fla	5-118
Bu.mal3	53724315	Y4A2W4MR2DB1-Y3Z	Fla	7-122
Bu.mal4	53724323	Y4A2W4MR2DB1-Y3Z	Fla	5-119
Bu.pse1	53720910	Y2A1W2MR1DB1Y1Z	Fla	18-133
Bu.pse2	53720917	Y2A1W2MR1DB1Y1Z	Fla	5-119
Bu.pse3	53721840	Y3C1	Fla	5-118
Bu.pse4	53723361	Y4C2	Fla	5-118
Bu.tha1	83716053	Y1C2	Fla	5-118
Bu.tha2	83716739	Y2C1	Fla	5-118
Bu.tha3	83717918	Y3A1W2MR1D1B2B1	Fla	4-118
Bu.tha4	83720536	Y4A3W4MR3D2B4Y5Z	Fla	5-119
Bu.tha5	83720689	Y4A3W4MR3D2B4Y5Z	Fla	7-122
Bu.xen1	91777132	M--Y1	Fla	4-119
Bu.xen2	91780059	Y2C	Fla	5-118
Bu.xen3	91785652	Y4AWMR4DB4Y3Z	Fla	7-122
Bu.xen4	91785659	Y4AWMR4DB4Y3Z	Fla	5-119
Burkh1	78060865	Y1C	Fla	5-118
Burkh2	78062824	M--Y2	Fla	4-119
Burkh3	78064834	Y3A2W3MR2DB3Y4Z	Fla	5-119
Burkh4	78064841	Y3A2W3MR2DB3Y4Z	Fla	7-122
Ca.cre1	16124687	M-M-Y1A1W1R1B1Y2D	Fla	5-119
Ca.cre2	16124692	M-M-Y1A1W1R1B1Y2D	Fla	9-124
Ca.cre3	16124845	r2b2y4w2a2m-Y3ZM	Fla	8-123
Ca.cre4	16124850	mzy3-MA2W2Y4B2R2	Fla	4-118
Ca.cre5	16127701	Y5R3	Fla	4-118
Ca.hyd1	78042666	DR2CY1----M-MY3-R1MW1B2Y2A2	Fla	5-118
Ca.hyd2	78043687	DR2CY1----M-MY3-R1MW1B2Y2A2	Fla	3-115
Ca.hyd3	78045144	DR2CY1----M-MY3-R1MW1B2Y2A2	Fla	4-118

Table A.11 (continued)

ID	GI	Gene Neighborhood	Class	Range
Ca.jej	15792443	Y	Fla	3-119
Ch.sal	92114140	AWMRBMYZ	Fla	8-123
Ch.vio1	34496471	Y1-A1MW1MD1B1	Fla	5-119
Ch.vio2	34497967	MY2-A2MW2R2B3	Fla	5-119
Ch.vio3	34498344	Y3	Fla	6-121
Ch.vio4	34498898	a4zy5v3v2--Y4A3W3--MR3B5D2	Fla	5-119
Ch.vio5	34498903	d2b5r3m--w3a3y4--V2V3Y5ZA4	Fla	8-123
Cl.ace1	15893413	Y1A1W1MR1Y2	Fla	5-118
Cl.ace2	15893418	Y1A1W1MR1Y2	Fla	5-118
Cl.ace3	15895486	W3DBR2A2CY3W2	Fla	4-117
Cl.tet	28211381	W2DBRACYW1	Fla	4-117
Co.psy1	71280820	Y1C1	Fla	1-114
Co.psy2	71282029	Y2ZAB---W2W1	Fla	7-122
De.aro1	71906364	MY1A1W1MMR1D1B1-V1V2Y2ZA2	Fla	5-119
De.aro2	71906375	MY1A1W1MMR1D1B1-V1V2Y2ZA2	Fla	9-124
De.aro3	71906778	Y3M-A3MW3D2R2D3B2	Fla	4-119
De.aro4	71907228	Y4	Fla	8-121
De.aro5	71909154	Y5	Fla	6-122
De.aro6	71909509	Y7Y6W4MA4	Tfp	6-120
De.aro7	71909510	Y7Y6W4MA4	Tfp	12-126
De.des1	78355431	Y1	Fla	9-124
De.des2	78355821	Y2	Fla	3-128
De.des3	78356619	B1-R2-W3Y3A1	Fla	5-121
De.des4	78356737	Y4	Fla	3-128
De.des5	78357150	MY5A2R3B2	Fla	5-118
De.geo	94985208	Y	Tfp	8-118
De.haf1	89895726	Y1	Fla	5-118
De.haf2	89896123	M--Y2	Fla	3-127
De.psy1	51245406	Y1C	Fla	3-118
De.psy2	51246503	Y2X2X1---r2b3w2a	Fla	5-122
De.rad	15808009	YWMMMA	Tfp	71-185
De.vul1	46579838	Y1	Fla	3-128
De.vul2	46580004	Y2A1R2B2	Fla	5-118
De.vul3	46580478	B3-R3-W4Y3A3	Fla	5-120
De.vul4	46581174	Y4	Fla	3-128
De.vul5	46581630	Y5	Fla	9-124
Er.car	50120628	m-----AWMRBYZ	Fla	8-123
Er.lit	85375078	AWYBR-M	Fla	4-118
Es.col	15802294	AWMMRBYZ	Fla	8-123
Ge.kau1	56419767	Y1	Fla	4-117
Ge.kau2	56419875	Y2	Fla	4-117
Ge.met1	78222817	Y1X2	Fla	4-119
Ge.met2	78223505	Y3A2-W3-W2-R3B3-Y2	Fla	8-123
Ge.met3	78223515	Y3A2-W3-W2-R3B3-Y2	Fla	4-119
Ge.met4	78223631	Y5--Y4A3--MMMW4R4D2B4	Fla	12-126
Ge.met5	78223634	Y5--Y4A3--MMMW4R4D2B4	Fla	3-128

Table A.11 (continued)

ID	GI	Gene Neighborhood	Class	Range
Ge.met6	78224311	Y6X4X3	Fla	4-119
Ge.met7	78224402	R7Y7CD3	Fla	6-119
Ge.sul1	39995511	MM-Y1X2X3	Fla	4-119
Ge.sul2	39996388	Y2M-Y3A2---M---MW4MW5-MM	Fla	3-128
Ge.sul3	39996391	Y2M-Y3A2---M---MW4MW5-MM	Fla	5-119
Ge.sul4	39996719	Y4X4	Fla	4-119
Ge.sul5	39997310	Y6A3-W7-W6--R4B3-Y5	Fla	6-121
Ge.sul6	39997321	Y6A3-W7-W6--R4B3-Y5	Fla	1-107
Ge.sul7	39998288	R5MW10Y7A4CD3	Fla	5-118
Gl.oxy1	58039984	M-Y1AWRBY2	Fla	3-117
Gl.oxy2	58039989	M-Y1AWRBY2	Fla	9-124
Ha.che1	83643354	M----Y1A1W1MMW2-R1D1B1	Fla	4-118
Ha.che2	83643440	Y3Y2W4MR2A2B2W3	Tfp	4-118
Ha.che3	83643441	Y3Y2W4MR2A2B2W3	Tfp	10-124
Ha.che4	83643562	Y4C1	Fla	5-119
Ha.che5	83646438	Y5-A3-MW5R3D2B3	Fla	28-142
Ha.che6	83647838	Y6ZA5B5---W9W8	Fla	7-122
Ha.che7	83648467	Y7Y8W10-MA6-W11	Tfp	120-234
Ha.che8	83648468	Y7Y8W10-MA6-W11	Tfp	5-119
Ha.mar	55378888	YC3D	Fla	4-116
Halob	15790091	W2YBAC2C1DR	Fla	5-117
He.aci	109947708	Y	Fla	3-119
He.hep	32266324	Y-----V2	Fla	3-119
He.pyl	15611426	Y	Fla	3-119
Id.loi	56460224	YZAB-W2W1	Fla	3-118
Janna1	89055329	db2-Y2AWR2Y1	Fla	9-124
Janna2	89055333	db2-Y2AWR2Y1	Fla	5-119
La.int1	94986969	Y1	Fla	9-124
La.int2	94987584	B-R-W1Y2AA	Fla	5-120
Le.int1	24213953	W1A1/CB1Y1	Fla	4-117
Le.int2	24215123	Y2-A2MW3D1B3	Fla	21-135
Li.inn	16799774	R-----V-YA	Fla	4-117
Li.mon	16802733	R-----V-YA	Fla	4-117
Ma.mag1	83309423	A1W1Y1B1R1	Fla	6-120
Ma.mag10	83313251	Y10C--M	Fla	6-119
Ma.mag2	83311734	Y2--Y3-Y4	Fla	5-118
Ma.mag3	83311737	Y2--Y3-Y4	Fla	6-119
Ma.mag4	83311739	Y2--Y3-Y4	Fla	4-118
Ma.mag5	83311747	M---Y5	Fla	21-136
Ma.mag6	83312037	b3/r3-----Y6	Fla	6-119
Ma.mag7	83312334	Y7	Tfp	12-127
Ma.mag8	83312796	Y8---M	Fla	12-124
Ma.mag9	83313244	Y9	Tfp	11-130
Me.ace1	20088915	w1m--Y1B1A1R1C1D1	Fla	14-127
Me.ace2	20091886	MW2-Y2B3A2C2D2R3	Fla	4-117
Me.bar	73668522	wm-YB1AR1D	Fla	4-117

Table A.11 (continued)

ID	GI	Gene Neighborhood	Class	Range
Me.bur	91772412	MW-YB1ACDR1	Fla	4-117
Me.fla1	91775595	Y1Y2W1MA1	Tfp	11-125
Me.fla2	91775596	Y1Y2W1MA1	Tfp	6-120
Me.fla3	91776281	Y4A2W3MMR1DB1Y3Z	Fla	8-123
Me.fla4	91776289	Y4A2W3MMR1DB1Y3Z	Fla	5-119
Me.hun1	88601427	Y1B1A1/C1Dc2	Fla	4-117
Me.hun2	88601444	Y2-R1	Fla	4-117
Me.hun3	88601626	Y3	Fla	4-117
Me.hun4	88602715	Y4	Fla	4-117
Me.hun5	88603217	Y5	Fla	4-116
Me.hun6	88604263	Y6----Y7Y8	Fla	4-117
Me.hun7	88604268	Y6----Y7Y8	Fla	4-117
Me.hun8	88604269	Y6----Y7Y8	Fla	3-116
Me.mar	45358496	YC2C1Rmdabw	Fla	7-121
Me.maz1	21226432	MW1-Y1B1A1C1D1R1	Fla	4-117
Me.maz2	21227429	w2m-Y2B2A2R2C2D2	Fla	24-137
Mo.the	83589655	DRCY	Fla	5-118
My.xan1	108757600	A7W1M-R6B2--Y1---b3a3mw3r2w5	Tfp	12-126
My.xan2	108758104	Y2	Fla	7-121
My.xan3	108758325	M-W13Y3A8C	Fla	5-118
My.xan4	108758942	W10R7MY4W8A6	Tfp	5-119
My.xan5	108759520	Y7A1W9W4--R8B1Y5	Fla	7-122
My.xan6	108759817	Y6-----W12	Fla	13-127
My.xan7	108761118	Y7A1W9W4--R8B1Y5	Fla	5-120
My.xan8	108763354	Y8W2R5W6A2B5M	Fla	8-124
My.xan9	108763514	Y9	Tfp	14-128
Na.pha1	76800910	Y1	Fla	5-117
Na.pha2	76801695	Y2C1D	Fla	3-115
Ni.eur1	30249010	Y1W1	Tfp	6-120
Ni.eur2	30249790	Y2	Tfp	13-127
Ni.eur3	30249875	Y3Z	Fla	8-123
Ni.ham1	92118626	Y1	Fla	9-124
Ni.ham2	92118835	AWY2B2R2	Fla	4-118
Ni.oce1	77163668	Y2Y1W2MR1AB1W1	Tfp	4-118
Ni.oce2	77163669	Y2Y1W2MR1AB1W1	Tfp	14-128
Ni.win1	75674720	AWY1BR-----M	Fla	4-118
Ni.win2	75676724	Y2	Fla	9-124
Nosto1	17228424	Y2Y1W1MA1	Tfp	4-118
Nosto2	17228425	Y2Y1W1MA1	Tfp	276-398
Nosto3	17228566	Y4Y3W2MA2	Tfp	4-118
Nosto4	17228567	Y4Y3W2MA2	Tfp	261-375
Nosto5	17229656	Y6Y5W3MA3	Tfp	9-123
Nosto6	17229657	Y6Y5W3MA3	Tfp	231-345
Nosto7	17231086	Y7	Fla	13-127
Oc.ihe	23099024	Y	Fla	5-118
Pe.car1	77918809	A2W4R1BD---Y1X2X3	Fla	5-122

Table A.11 (continued)

ID	GI	Gene Neighborhood	Class	Range
Pe.car2	77919012	A3--Y2	Fla	6-121
Ph.lum	37525786	AWMMRBYZ	Fla	8-123
Ph.pro1	54302440	V1Y1	Fla	39-161
Ph.pro2	54307967	MY2-A1-MR1B1	Fla	5-119
Ph.pro3	54308133	Y3ZA2B2-W1W2	Fla	7-122
Polar1	91787034	Y1Y2W1MA1	Tfp	10-124
Polar2	91787035	Y1Y2W1MA1	Tfp	6-120
Polar3	91789131	Y3X	Fla	3-119
Ps.aer1	15595377	MY1A1W1MR1DB1	Fla	5-119
Ps.aer2	15595605	Y2Y3W2MR2A2B2W3	Tfp	10-124
Ps.aer3	15595606	Y2Y3W2MR2A2B2W3	Tfp	4-118
Ps.aer4	15596653	Y4ZA3B3---W4W5	Fla	3-118
Ps.arc1	71066372	Y2Y1WMRA	Tfp	4-118
Ps.arc2	71066373	Y2Y1WMRA	Tfp	10-124
Ps.atl1	109899334	Y1ZAB-W2W1	Fla	7-122
Ps.atl2	109899569	Y2	Fla	7-120
Ps.cry1	93005216	Y1X1X2	Fla	4-117
Ps.cry2	93006924	Y3Y2WMRAB	Tfp	4-118
Ps.cry3	93006925	Y3Y2WMRAB	Tfp	10-124
Ps.ent1	104782800	Y1ZA2B3---W4W3	Fla	3-118
Ps.ent2	104783972	Y3Y2W6MA3W5	Tfp	4-118
Ps.ent3	104783973	Y3Y2W6MA3W5	Tfp	1-113
Ps.flu1	70729058	Y1ZA2B2---W3W4	Fla	3-118
Ps.flu2	70733111	Y3Y2W6MA3W5	Tfp	4-118
Ps.flu3	70733112	Y3Y2W6MA3W5	Tfp	10-124
Ps.hal1	77359759	Y1ZAB---W1W2	Fla	7-122
Ps.hal2	77359861	Y2	Fla	7-120
Ps.hal3	77360088	Y3	Tfp	13-126
Ps.hal4	77360443	Y4	Fla	10-123
Ps.hal5	77361296	Y5	Fla	6-119
Ps.put1	26991030	Y1ZA2B3---W4W3	Fla	3-118
Ps.put2	26991668	Y3Y2W6MA3W5	Tfp	4-118
Ps.put3	26991669	Y3Y2W6MA3W5	Tfp	7-121
Ps.syr1	28868135	MY1-A1MW1R1DB1	Fla	5-119
Ps.syr2	28869184	Y2ZA3B3---W4W5	Fla	3-118
Ps.syr3	28872147	Y4Y3W8MA4W7	Tfp	4-118
Ps.syr4	28872148	Y4Y3W8MA4W7	Tfp	10-124
Py.aby	14521754	wmRYBAC2C1DM	Fla	4-117
Py.hor	14590394	wm-RYBAC1C2DM	Fla	4-117
Ra.eut1	73539421	M-Y1W3MM-----A2W4R4DB4Y2Z	Fla	9-123
Ra.eut2	73539435	M-Y1W3MM-----A2W4R4DB4Y2Z	Fla	7-122
Ra.eut3	73541643	M-Y3	Fla	4-119
Ra.eut4	73542307	Y5Y4W5MA3	Tfp	7-121
Ra.eut5	73542308	Y5Y4W5MA3	Tfp	34-148
Ra.met1	94309614	Y1Y2W1MA1	Tfp	39-153
Ra.met2	94309615	Y1Y2W1MA1	Tfp	6-120

Table A.11 (continued)

ID	GI	Gene Neighborhood	Class	Range
Ra.met3	94312612	M-Y3W2MM-----A2W3R2DB1Y4Z	Fla	6-120
Ra.met4	94312626	M-Y3W2MM-----A2W3R2DB1Y4Z	Fla	18-133
Ra.sol1	17545387	Y1Y2W1MA1	Tfp	56-170
Ra.sol2	17545388	Y1Y2W1MA1	Tfp	6-120
Ra.sol3	17545460	Y3Z1	Fla	8-125
Ra.sol4	17549621	Y5A2W2MR1DBY4Z2	Fla	8-125
Ra.sol5	17549628	Y5A2W2MR1DBY4Z2	Fla	6-123
Rh.etl1	86356289	M-Y1A1W1R1B1Y2D	Fla	5-119
Rh.etl2	86356294	M-Y1A1W1R1B1Y2D	Fla	9-124
Rh.etl3	86358273	Y3	Tfp	4-119
Rh.etl4	86359114	Y4A2W4MMMW3R2B2	Fla	5-119
Rh.fer1	89899377	Y1A1W2R1D1B1	Fla	4-118
Rh.fer2	89899707	MY2-A2MW3R2D2B2	Fla	5-119
Rh.fer3	89900165	Y3Y4W6MA3	Tfp	10-124
Rh.fer4	89900166	Y3Y4W6MA3	Tfp	6-120
Rh.fer5	89900675	Y5C	Fla	38-151
Rh.fer6	89902464	Y6Z	Fla	7-122
Rh.pal1	39933220	Y1A1W2W1MR1B1	Fla	4-118
Rh.pal2	39934250	Y2	Fla	9-124
Rh.pal3	39934699	A2W3Y3B2R2	Fla	4-118
Rh.rub1	83591861	A1Y1B1R1	Fla	4-118
Rh.rub2	83592734	Y2A2W1MW2MMR2B2Dm	Fla	6-120
Rh.rub3	83593661	MW3B4-R4Y3A3--Y4--M	Fla	13-127
Rh.rub4	83593665	MW3B4-R4Y3A3--Y4--M	Fla	5-120
Rh.rub5	83594169	Y5Z3	Fla	9-125
Rh.sph1	77462123	Y1A1W1W2R1B1	Fla	5-119
Rh.sph2	77462991	Y4M-MD-Y3A2W3R3Y2	Fla	9-124
Rh.sph3	77462995	Y4M-MD-Y3A2W3R3Y2	Fla	5-119
Rh.sph4	77463001	Y4M-MD-Y3A2W3R3Y2	Fla	5-119
Rh.sph5	77463614	A3R4B3W4-MY5A4	Fla	5-131
Rh.sph6	77465308	Y6M	Fla	5-119
Sa.deg1	90020171	Y2Y1W1MA1	Tfp	6-120
Sa.deg2	90020172	Y2Y1W1MA1	Tfp	190-304
Sa.deg3	90020549	Y3	Fla	5-118
Sa.deg4	90021808	Y4ZA2B1---W3W2	Fla	7-122
Sa.deg5	90022750	mm--Y5A3W4MR3DB3	Fla	4-119
Sa.deg6	90023273	Y7Y6W6MR4A4B4W5	Tfp	4-118
Sa.deg7	90023274	Y7Y6W6MR4A4B4W5	Tfp	10-124
Sa.ent	16760865	AWMRBYZ	Fla	8-123
Sa.rub	83816239	WR2Y-AB1-M-M-M	Fla	13-128
Sa.typ	16765258	Y	Fla	8-123
Sh.boy	82543638	RBYZ	Fla	8-123
Sh.den1	91792702	Y1ZA1B1--W1W2	Fla	7-122
Sh.den2	91794642	MY2-A2MW3R2DB2	Fla	5-119
Sh.fle	30063334	AWMMRBYZ	Fla	8-123
Sh.one1	24373680	M--Y1A1W1MR1D1B1	Fla	5-119

Table A.11 (continued)

ID	GI	Gene Neighborhood	Class	Range
Sh.one2	24373867	Y2--MW2R2D2B2	Fla	4-118
Sh.one3	24374090	Y3C2	Fla	5-118
Sh.one4	24374721	Y4ZA2B3--W4W3	Fla	7-122
Sh.son	74311768	AWMMRBYZ	Fla	8-123
Si.mel1	15964391	M-Y1A1W1R1B1Y2D	Fla	5-119
Si.mel2	15964396	M-Y1A1W1R1B1Y2D	Fla	9-124
Silic1	99078181	d1b1m-Y2AW1R1Y1	Fla	9-124
Silic2	99078185	d1b1m-Y2AW1R1Y1	Fla	5-119
Sp.ala	103487223	AWYBR	Fla	5-121
Sy.aci1	85859036	m--w3---MR2D1-Y1-D2B2A2Y2X1	Fla	7-124
Sy.aci2	85859041	m--w3---MR2D1-Y1-D2B2A2Y2X1	Fla	7-122
Sy.elo1	56750338	Y1	Tfp	4-118
Sy.elo2	56750693	Y3Y2W3MA2W2	Tfp	4-118
Sy.elo3	56750694	Y3Y2W3MA2W2	Tfp	300-414
Sy.the	51892678	W2W3ACDYBRX1	Fla	4-116
Synco1	86605197	Y1	Tfp	5-119
Synco2	86606801	Y3Y2W1---MA1	Tfp	4-118
Synco3	86606802	Y3Y2W1---MA1	Tfp	298-412
Synco4	86607341	Y4Y5Y6W2MA2	Tfp	255-369
Synco5	86607342	Y4Y5Y6W2MA2	Tfp	4-118
Synco6	86607343	Y4Y5Y6W2MA2	Tfp	4-118
Syncy1	16329617	Y1Y2W1M	Tfp	268-384
Syncy2	16329618	Y1Y2W1M	Tfp	30-144
Syncy3	16329793	Y4Y3W2MA1	Tfp	4-119
Syncy4	16329794	Y4Y3W2MA1	Tfp	282-397
Syncy5	16331990	Y6Y5W4MMA3W3	Tfp	4-118
Syncy6	16331991	Y6Y5W4MMA3W3	Tfp	278-392
Tb.den1	74317632	Y2A1W1M-MMMRDBY1Z	Fla	8-123
Tb.den2	74317643	Y2A1W1M-MMMRDBY1Z	Fla	4-118
Tb.den3	74318569	Y4Y3W2MA2	Tfp	6-120
Tb.den4	74318570	Y4Y3W2MA2	Tfp	11-125
Th.elo1	22297888	Y1Y2W1MA1	Tfp	253-367
Th.elo2	22297889	Y1Y2W1MA1	Tfp	4-118
Th.elo3	22298114	Y4Y3W3MA2W2	Tfp	4-118
Th.elo4	22298115	Y4Y3W3MA2W2	Tfp	319-433
Th.elo5	22298568	Y6Y5W4MA3	Tfp	4-119
Th.elo6	22298569	Y6Y5W4MA3	Tfp	257-371
Th.kod	57640567	wmRYBAAC1C2MD	Fla	4-117
Th.mar	15643463	AW1Y	Fla	5-118
Th.ten1	20807517	MW2B1-R1Y1A1	Fla	16-130
Th.ten2	20807875	Y2	Fla	4-117
Tm.cru1	78485093	Y1ZA1W2--y2w3-RBM	Fla	9-124
Tm.cru2	78485099	mbr-W3Y2--w2a1zy1	Fla	4-119
Tm.cru3	78485439	Y3	Fla	4-117
Tm.den1	78777168	Y1W1M-A1R1-DB1-Z1M-M----mM	Fla	5-125
Tm.den2	78777232	Y2C	Fla	3-117

Table A.11 (continued)

ID	GI	Gene Neighborhood	Class	Range
Tm.den3	78777753	Y3	Fla	4-117
Tm.den4	78777869	Y4	Fla	3-119
Tr.den	42527002	AW1/R2XY	Fla	27-142
Tr.pal	15639357	AW1/R1XY	Fla	27-142
Vi.cho1	15601845	Y1A1W2W1MR1DM	Fla	10-124
Vi.cho2	15641095	Y2C2	Fla	5-118
Vi.cho3	15641409	mm-MW3B1-R2Y3A2-M	Fla	5-119
Vi.cho4	15642065	Y4ZA3B2-W5W4	Fla	11-126
Vi.fis	59712440	YZAB--W2W1	Fla	3-118
Vi.par	28899005	YZAB-W2W1	Fla	7-122
Vi.vul1	27365298	Y1ZA1B1-W1W2	Fla	3-118
Vi.vul2	27367550	Y2A2W4W3MR3DB3M	Fla	4-118
Wo.suc1	34557035	Y1	Fla	3-119
Wo.suc2	34557459	Y2W2	Fla	8-128
Xa.axo1	21241908	Y1	Tfp	4-118
Xa.axo2	21242648	W5-Y2A1M-MM-MMMMMMMR2DB2	Fla	5-119
Xa.axo3	21242676	Y3ZA4	Fla	3-118
Xa.axo4	21242712	Y4	Fla	6-120
Xa.axo5	21243828	Y6Y5W4MA3B4W3	Tfp	4-118
Xa.axo6	21243829	Y6Y5W4MA3B4W3	Tfp	16-130
Xa.cam1	21230507	Y1	Tfp	4-118
Xa.cam2	21231334	W6-Y2A1M--MMMM-MM-M-W1-MR2DB2	Fla	5-119
Xa.cam3	21231353	Y3ZA2	Fla	3-118
Xa.cam4	21231382	Y4	Fla	6-120
Xa.cam5	21232355	Y6Y5W5MA4B5W4	Tfp	4-118
Xa.cam6	21232356	Y6Y5W5MA4B5W4	Tfp	16-130
Xa.ory1	58580534	Y1	Tfp	4-118
Xa.ory2	58581370	Y6Y2W2MA2A2B2W3	Tfp	4-118
Xa.ory3	58582211	Y3	Fla	6-120
Xa.ory4	58582245	Y4ZA3	Fla	7-122
Xa.ory5	58582458	W4-Y5A4-MM-M-MMMM-W5----R2DB3	Fla	5-119
Xa.ory6	77760636	Y6Y2W2MA2A2B2W3	Tfp	16-130
Xy.fas1	15837052	Y1	Tfp	8-122
Xy.fas2	77747573	Y2W2MABW1	Tfp	16-130
Ye.pes	16121942	MRBYZ	Fla	8-123
Ye.pse	51596720	m-----AW--MMRBYZ	Fla	8-123
Zy.mob	56550975	M-A1RB1DYW	Fla	5-119

Table A.12 40H class MCP data. ID, GI, Gene Neighborhood, and Range are explained in Table A.2. Class is based on phylogenetic analysis of a 40H multiple alignment. A minimum evolution tree was built from the 40H MCP alignment in MEGA with pairwise deletions and the pairwise distance.

ID	GI	Gene Neighborhood	Class	Range
An.deh	86157040	b1r1w1a1d1-y2--y3W2W3MA2B2R2	Alt	138-418
An.var1	75906286	Y1Y2W1MA1	Tfp	471-751
An.var2	75906729	Y4Y3W2MA2	Tfp	696-976
An.var3	75909982	Y6Y7W3MA3	Tfp	819-1099
Azoar1	56476384	Y1Y2WMA	Tfp	433-713
Azoar2	56478990	M	F6-mix	388-668
Bu.cen	107026920	MW3R2-W2A2B3	Alt	276-556
Bu.mal	53716498	MW2R1W1A1B2	Alt	276-556
Bu.pse	53722895	MW4R2W3A2B2	Alt	276-556
Bu.tha	83716541	MW1R2W3A2B3	Alt	276-556
Burkh	78063155	MW2R1W1A1B2	Alt	276-556
Ch.vio1	34495486	M	F6-beta	380-660
Ch.vio2	34495550	M	F6-beta	262-542
Ch.vio3	34495712	M	F6-beta	260-540
Ch.vio4	34495720	M	F6-beta	262-542
Ch.vio5	34496082	M	F6-beta	263-543
Ch.vio6	34496168	M	F6-beta	410-690
Ch.vio7	34496328	M	F6-mix	117-397
Ch.vio8	34496354	M	F6-beta	383-663
Ch.vio9	34496535	MM	F6-beta	348-628
Ch.vio10	34496536	MM	F6-beta	347-627
Ch.vio11	34496783	M	F6-beta	252-532
Ch.vio12	34496813	M	F6-mix	146-426
Ch.vio13	34496872	M	F6-beta	268-548
Ch.vio14	34496979	M		79-359
Ch.vio15	34497171	M	F6-beta	413-693
Ch.vio16	34497305	M	F6-beta	262-542
Ch.vio17	34497918	M	F6-beta	264-544
Ch.vio18	34498307	M	F6-beta	222-502
Ch.vio19	34498314	M	F6-beta	258-538
Ch.vio20	34498362	M	F6-beta	255-535
Ch.vio21	34498740	M	F6-gamma	348-628
Ch.vio22	34498775	MM	F6-beta	378-658
Ch.vio23	34499104	M	F6-beta	391-671
Ch.vio24	34499321	M	F6-beta	262-542
Ch.vio25	34499610	MM	F6-beta	265-545
Ch.vio26	34499611	MM	F6-beta	345-625
Ch.vio27	34499699	M	F6-beta	268-548
Ch.vio28	34499752	M	F6-beta	265-545
Co.psy1	71277838	M	F6-gamma	392-672
Co.psy2	71278287	M	F6-gamma	287-567
Co.psy3	71278459	M	F6-gamma	78-358
Co.psy4	71278766	M	F6-gamma	360-640

Table A.12 (continued)

ID	GI	Gene Neighborhood	Class	Range
Co.psy5	71278893	M	F6-gamma	1-281
Co.psy6	71279028	M---M	F6-gamma	223-503
Co.psy7	71279083	M--M	F6-gamma	1-281
Co.psy8	71279453	M	F6-gamma	424-704
Co.psy9	71279484	M	F6-gamma	353-633
Co.psy10	71279686	M	F6-gamma	231-511
Co.psy11	71279880	M	F6-gamma	355-635
Co.psy12	71280322	M	F6-gamma	1-281
Co.psy13	71280537	M	F6-gamma	253-533
Co.psy14	71280713	M	F6-gamma	255-535
Co.psy15	71280779	M---M	F6-gamma	394-674
Co.psy16	71280933	M	F6-gamma	381-661
Co.psy17	71281197	M	F6-gamma	421-701
Co.psy18	71281232	M--M	F6-gamma	264-544
Co.psy19	71281444	M	F6-gamma	381-661
Co.psy20	71281933	M	F6-gamma	205-485
Co.psy21	71282390	M	F6-gamma	339-619
Co.psy22	71282609	M	F6-gamma	277-557
De.aro1	71906201	M	F6-mix	55-335
De.aro2	71906313	M	F6-beta	264-544
De.aro3	71906476	M	F6-beta	400-680
De.aro4	71906565	M	F6-beta	259-539
De.aro5	71906995	M	F6-beta	264-544
De.aro6	71907531	M	F6-beta	267-547
De.aro7	71908341	M	F6-beta	264-544
De.aro8	71908447	M	F6-beta	391-671
De.aro9	71908456	M	F6-beta	116-396
De.aro10	71908496	M	F6-beta	257-537
De.aro11	71908546	M	F6-beta	260-540
De.aro12	71908710	x----M	F6-mix	150-430
De.aro13	71908725	M	F6-beta	261-541
De.aro14	71909191	M	F6-beta	274-554
De.aro15	71909422	M	F6-beta	250-530
De.aro16	71909507	Y7Y6W4MA4	Tfp	425-705
De.des1	78355265	M	F4a	365-645
De.des2	78355394	M	F4a	332-612
De.des3	78355416	M	F4a	327-607
De.des4	78355481	M	F4a	153-433
De.des5	78355505	M	F4a	441-721
De.des6	78355750	M	F4a	1-281
De.des7	78356326	M	F4a	446-726
De.des8	78356369	M	F4a	395-675
De.des9	78356693	M	F4a	531-811
De.des10	78356708	M	F4a	491-771
De.des11	78356795	M--m	F4a	431-711

Table A.12 (continued)

ID	GI	Gene Neighborhood	Class	Range
De.des12	78357454	M	F4a	526-806
De.des13	78357627	M	F4a	267-547
De.des14	78357776	M	F4a	471-751
De.des15	78357855	M	F4a	1-281
De.des16	78357898	M	F4a	394-674
De.des17	78358008	M	F4a	392-672
De.des18	78358060	M	F4a	416-696
De.des19	78358114	M-----v3	F4a	1-281
De.des20	78358286	M	F4a	321-601
De.des21	78358301	M	F4a	332-612
De.des22	78358333	M-M	F4a	438-718
De.des23	78358335	M-M	F4a	438-718
De.des24	78358547	M	F4a	409-689
De.des25	78358600	M	F4a	392-672
De.psy	51244830	M-M-----mM	F6-gamma	359-639
De.rad1	15808011	YWMMMA	Tfp	462-742
De.rad2	15808012	YWMMMA	Tfp	474-754
De.rad3	15808013	YWMMMA	Tfp	472-752
De.vul1	46578435	M	F4a	466-746
De.vul2	46578511	M	F4a	439-719
De.vul3	46578587	M	F4a	263-543
De.vul4	46578600	M	F4a	306-586
De.vul5	46578760	M	F4a	358-638
De.vul6	46579022	M	F4a	396-676
De.vul7	46579059	M	F4a	398-678
De.vul8	46579082	M	F4a	248-528
De.vul9	46579113	M	F4a	534-814
De.vul10	46579163	M	F4a	1-281
De.vul11	46579348	M	F4a	1-281
De.vul12	46579580	M	F4a	524-804
De.vul13	46579811	M	F4a	349-629
De.vul14	46580266	M	F4a	412-692
De.vul15	46580278	M	F4a	490-770
De.vul16	46580293	M	F4a	520-800
De.vul17	46580384	M	F4a	323-603
De.vul18	46580700	M	F4a	496-776
De.vul19	46580714	M	F4a	413-693
De.vul20	46580722	M	F4a	307-587
De.vul21	46580989	M	F4a	345-625
De.vul22	46581142	M	F4a	515-795
De.vul23	46581438	M	F4a	268-548
De.vul24	46581485	M	F4a	303-583
De.vul25	46581557	M	F4a	536-816
De.vul26	46581584	M	F4a	389-669
Ge.met1	78221750	M	F4b	262-542
Ge.met2	78221932	M	F4b	268-548

Table A.12 (continued)

ID	GI	Gene Neighborhood	Class	Range
Ge.met3	78222019	M	F4b	108-388
Ge.met4	78222041	M	F4b	265-545
Ge.met5	78223905	W5R6W6MA4B6	Alt	257-537
Ge.met6	78224021	M-B7M	F4b	290-570
Ge.met7	78224024	M-B7M	F4b	269-549
Ge.met8	78224135	Mm	F4b	245-525
Ge.met9	78224279	M	F4b	527-807
Ge.sul1	39995508	MM-Y1X2X3	F4b	267-547
Ge.sul2	39995509	MM-Y1X2X3	F4b	257-537
Ge.sul3	39995689	Mm	F4b	245-525
Ge.sul4	39995856	m-----M	F4b	259-539
Ge.sul5	39995862	m-----M	F4b	268-548
Ge.sul6	39996019	M	F4b	112-392
Ge.sul7	39996038	M	F4b	246-526
Ge.sul8	39996132	MM-MMM-----M	F4b	268-548
Ge.sul9	39996133	MM-MMM-----M	F4b	268-548
Ge.sul10	39996135	MM-MMM-----M	F4b	241-521
Ge.sul11	39996136	MM-MMM-----M	F4b	253-533
Ge.sul12	39996137	MM-MMM-----M	F4b	267-547
Ge.sul13	39996143	MM-MMM-----M	F4b	259-539
Ge.sul14	39996476	M	F4b	260-540
Ge.sul15	39997674	MW9	F4b	262-542
Ge.sul16	39997746	M	F4b	112-392
Ge.sul17	39998033	M	F4b	262-542
Ha.che1	83642995	M	F6-gamma	261-541
Ha.che2	83643409	M	Alt	343-623
Ha.che3	83643438	Y3Y2W4MR2A2B2W3	Tfp	391-671
Ha.che4	83644002	M	F6-gamma	386-666
Ha.che5	83644171	M	Alt	267-547
Ha.che6	83644493	M	F6-gamma	301-581
Ha.che7	83644586	M	F6-gamma	233-513
Ha.che8	83644748	M	F6-gamma	396-676
Ha.che9	83644758	M	F6-gamma	349-629
Ha.che10	83644767	M	F6-gamma	303-583
Ha.che11	83645255	M	F6-gamma	284-564
Ha.che12	83645370	M	F6-gamma	403-683
Ha.che13	83645863	M	F6-gamma	244-524
Ha.che14	83646110	M	F6-gamma	265-545
Ha.che15	83646199	M	F6-gamma	284-564
Ha.che16	83646471	M	F6-gamma	382-662
Ha.che17	83647021	M	F6-gamma	268-548
Ha.che18	83647345	m--M	F6-gamma	20-300
Ha.che19	83647533	M	F6-gamma	373-653
Ha.che20	83647556	M	F6-beta	392-672
Ha.che21	83647634	M	F6-gamma	368-648
Ha.che22	83647717	M	F6-gamma	210-490

Table A.12 (continued)

ID	GI	Gene Neighborhood	Class	Range
Ha.che23	83648708	M	F6-gamma	246-526
Ha.che24	83648937	M	F6-gamma	261-541
Ha.che25	83649336	M	F6-gamma	367-647
He.aci	109947957	M	F3	285-565
He.hep1	32266012	M	F3	252-532
He.hep2	32266654	M		276-556
He.hep3	32266809	M	F3	388-668
He.pyl	15611165	M---m	F3	284-564
Id.loi1	56459144	M	F6-gamma	241-521
Id.loi2	56459286	M	F6-gamma	406-686
Id.loi3	56459313	M	F6-gamma	349-629
Id.loi4	56459719	M	F6-gamma	261-541
Id.loi5	56460176	M	F6-gamma	262-542
Id.loi6	56460404	M	F6-gamma	265-545
Id.loi7	56460529	M	F6-gamma	382-662
Id.loi8	56460876	M	F6-gamma	332-612
Id.loi9	56460980	M	F6-gamma	282-562
Id.loi10	56461335	M	F6-gamma	386-666
Id.loi11	56461443	M	F6-gamma	265-545
La.int1	94972501	M----M	F4a	415-695
La.int2	94987285	M	F4a	301-581
Ma.mag	83312100	MB4R4A2--W4	Alt	376-656
Me.cap	53804249	M	Alt	341-621
Me.fla	91775598	Y1Y2W1MA1	Tfp	262-542
Me.hun	88602281	W7R3W6MA3B4	Alt	1-281
Me.lot	13488383	MW2RW1AB	Alt	263-543
My.xan1	108757399	W10R7MY4W8A6	Alt	258-538
My.xan2	108758729	W5R2W3MA3B3---y1--b2r6-mw1a7	Alt	267-547
My.xan3	108762109	W7-MMA5-B6R4	Alt	309-589
My.xan4	108762134	M	Alt	250-530
Ni.eur	30249233	MA1	Tfp	406-686
Ni.mul	82701466	W1RW2MAB	Alt	206-486
Ni.oce	77163666	Y2Y1W2MR1AB1W1	Tfp	255-535
Nosto1	17228422	Y2Y1W1MA1	Tfp	694-974
Nosto2	17228564	Y4Y3W2MA2	Tfp	821-1101
Nosto3	17229654	Y6Y5W3MA3	Tfp	471-751
Pe.car	77917670	M	F6-mix	99-379
Ph.pro1	54301708	M	F6-gamma	410-690
Ph.pro2	54301780	M	F6-gamma	384-664
Ph.pro3	54301884	M	F6-gamma	354-634
Ph.pro4	54301900	M	F6-gamma	1-281
Ph.pro5	54302045	M-m	F6-gamma	276-556
Ph.pro6	54302047	M-m	F6-gamma	260-540
Ph.pro7	54302061	M	F6-gamma	356-636
Ph.pro8	54302087	M	F6-gamma	348-628
Ph.pro9	54302188	mM	F6-gamma	237-517

Table A.12 (continued)

ID	GI	Gene Neighborhood	Class	Range
Ph.pro10	54302189	mM	F6-gamma	267-547
Ph.pro11	54302333	M	F6-gamma	248-528
Ph.pro12	54303005	M	F6-gamma	234-514
Ph.pro13	54303417	M	F6-gamma	297-577
Ph.pro14	54303523	M-Mm	F6-gamma	187-467
Ph.pro15	54303525	M-Mm	F6-gamma	385-665
Ph.pro16	54303526	Mm-m	F6-gamma	296-576
Ph.pro17	54303603	M	F6-gamma	264-544
Ph.pro18	54307660	M	F6-gamma	253-533
Ph.pro19	54308182	M	F6-gamma	371-651
Ph.pro20	54308329	M	F6-gamma	431-711
Ph.pro21	54308447	M	F6-gamma	360-640
Ph.pro22	54308562	M	F6-gamma	366-646
Ph.pro23	54308588	M	F6-gamma	262-542
Ph.pro24	54308659	M	F6-gamma	342-622
Ph.pro25	54309188	M	F6-gamma	239-519
Ph.pro26	54309203	M	F6-gamma	417-697
Ph.pro27	54309261	M	F6-gamma	386-666
Ph.pro28	54309325	M	F6-gamma	281-561
Ph.pro29	54309371	M	F6-gamma	186-466
Ph.pro30	54309471	M	F6-gamma	268-548
Ph.pro31	54309570	M	F6-gamma	387-667
Ph.pro32	54309987	MM	F6-gamma	346-626
Ph.pro33	54309988	MM	F6-gamma	347-627
Ph.pro34	54310281	M	F6-gamma	358-638
Ph.pro35	54310442	M	F6-gamma	347-627
Ph.pro36	54310451	M		258-538
Polar1	91787037	Y1Y2W1MA1	Tfp	411-691
Polar2	91788328	W2RW3--MMMA2B	Alt	232-512
Polar3	91788329	W2RW3--MMMA2B	Alt	296-576
Polar4	91788330	W2RW3--MMMA2B	Alt	275-555
Ps.aer1	15595608	Y2Y3W2MR2A2B2W3	Tfp	395-675
Ps.aer2	15596448	M	F6-gamma	261-541
Ps.aer3	15596758	M	F6-gamma	241-521
Ps.aer4	15596805	M	F6-gamma	261-541
Ps.aer5	15596843	M	F6-gamma	372-652
Ps.aer6	15597757	M	F6-gamma	288-568
Ps.aer7	15597769	M	F6-beta	255-535
Ps.aer8	15597848	m-M	F6-gamma	281-561
Ps.aer9	15597850	m-M	F6-gamma	434-714
Ps.aer10	15597984	M	F6-gamma	251-531
Ps.aer11	15598063	M	F6-gamma	210-490
Ps.aer12	15598116	M	F6-gamma	265-545
Ps.aer13	15598903	MW7R4W6A4B4	Alt	262-542
Ps.aer14	15599503	M-MM	F6-gamma	352-632
Ps.aer15	15599505	M-MM	F6-gamma	349-629

Table A.12 (continued)

ID	GI	Gene Neighborhood	Class	Range
Ps.aer16	15599506	M-MM	F6-gamma	349-629
Ps.aer17	15599716	M	F6-gamma	393-673
Ps.aer18	15599829	M	F6-gamma	432-712
Ps.aer19	15600037	M	F6-gamma	377-657
Ps.aer20	15600108	M	F6-gamma	261-541
Ps.aer21	15600265	M	F6-gamma	367-647
Ps.arc	71066370	Y2Y1WMRA	Tfp	168-448
Ps.atl1	109896399	M	F6-gamma	366-646
Ps.atl2	109896474	M	F6-gamma	259-539
Ps.atl3	109896848	M	F6-gamma	281-561
Ps.atl4	109896932	M	F6-gamma	358-638
Ps.atl5	109898564	M	F6-gamma	1-281
Ps.atl6	109899729	M	F6-gamma	394-674
Ps.atl7	109899765	M	F6-gamma	245-525
Ps.atl8	109900479	M-M	F6-gamma	218-498
Ps.atl9	109900481	M-M	F6-gamma	201-481
Ps.cry	93006922	Y3Y2WMRAB	Tfp	166-446
Ps.ent1	104779920	M	F6-gamma	368-648
Ps.ent2	104779960	M	F6-gamma	359-639
Ps.ent3	104780126	M	F6-gamma	433-713
Ps.ent4	104780151	M	F6-gamma	390-670
Ps.ent5	104780168	M---v1	F6-gamma	244-524
Ps.ent6	104780365	M	F6-gamma	344-624
Ps.ent7	104780451	MW1R2W2A1B1	Alt	260-540
Ps.ent8	104780695	M	F6-gamma	365-645
Ps.ent9	104780988	M	F6-gamma	348-628
Ps.ent10	104780998	M	F6-gamma	241-521
Ps.ent11	104781023	M	F6-gamma	212-492
Ps.ent12	104781177	M	F6-gamma	261-541
Ps.ent13	104781861	M	F6-beta	255-535
Ps.ent14	104782002	M	F6-gamma	264-544
Ps.ent15	104782061	M	F6-gamma	435-715
Ps.ent16	104782647	M	F6-gamma	261-541
Ps.ent17	104782755	M	F6-gamma	258-538
Ps.ent18	104782763	M	F6-gamma	241-521
Ps.ent19	104782926	M	F6-gamma	241-521
Ps.ent20	104783055	M	F6-gamma	408-688
Ps.ent21	104783791	M	F6-gamma	373-653
Ps.ent22	104783866	M	F6-gamma	378-658
Ps.ent23	104783970	Y3Y2W6MA3W5	Tfp	404-684
Ps.ent24	104784000	MM	F6-gamma	358-638
Ps.ent25	104784001	MM	F6-gamma	367-647
Ps.ent26	104784019	M	F6-gamma	369-649
Ps.ent27	104784077	m--M	F6-gamma	365-645
Ps.ent28	104784080	m--M	F6-gamma	261-541
Ps.ent29	104784305	M	F6-beta	375-655

Table A.12 (continued)

ID	GI	Gene Neighborhood	Class	Range
Ps.flu1	70728511	MW1R1W2A1B1	Alt	260-540
Ps.flu2	70729125	M	F6-gamma	261-541
Ps.flu3	70729147	M	F6-gamma	376-656
Ps.flu4	70729283	M	F6-gamma	241-521
Ps.flu5	70729309	M	F6-gamma	241-521
Ps.flu6	70729551	M	F6-gamma	261-541
Ps.flu7	70729680	M	F6-gamma	1-281
Ps.flu8	70729696	M	F6-beta	263-543
Ps.flu9	70729775	M	F6-gamma	306-586
Ps.flu10	70729905	M	F6-gamma	260-540
Ps.flu11	70730262	M	F6-gamma	383-663
Ps.flu12	70730480	M	F6-gamma	265-545
Ps.flu13	70730683	M	F6-gamma	241-521
Ps.flu14	70730719	M	F6-gamma	281-561
Ps.flu15	70731367	m-M	F6-gamma	279-559
Ps.flu16	70731369	m-M	F6-gamma	482-762
Ps.flu17	70731638	M	F6-gamma	213-493
Ps.flu18	70731886	M	F6-gamma	303-583
Ps.flu19	70731972	M	F6-gamma	446-726
Ps.flu20	70732109	M	F6-gamma	349-629
Ps.flu21	70732369	M---v3	F6-gamma	184-464
Ps.flu22	70732442	M	F6-gamma	433-713
Ps.flu23	70732694	M	F6-gamma	260-540
Ps.flu24	70733109	Y3Y2W6MA3W5	Tfp	395-675
Ps.flu25	70733232	M	F6-beta	255-535
Ps.flu26	70733311	M	F6-beta	1-281
Ps.flu27	70733635	M	F6-gamma	346-626
Ps.flu28	70733899	M	F6-gamma	389-669
Ps.flu29	70733922	M	F6-gamma	358-638
Ps.flu30	70733993	m--MM	F6-gamma	265-545
Ps.flu31	70733994	m--MM	F6-gamma	285-565
Ps.flu32	70733997	mm--M	F6-gamma	412-692
Ps.flu33	70734076	M	F6-gamma	364-644
Ps.flu34	70734176	M	F6-gamma	373-653
Ps.flu35	70734209	M	F6-gamma	368-648
Ps.flu36	70734275	M	F6-gamma	359-639
Ps.flu37	70734347	M	F6-gamma	385-665
Ps.hal1	77359361	M	F6-gamma	232-512
Ps.hal2	77359662	M	F6-gamma	275-555
Ps.hal3	77359797	M	F6-gamma	357-637
Ps.hal4	77359883	M	F6-gamma	235-515
Ps.hal5	77361303	M	F6-gamma	264-544
Ps.hal6	77361553	M	F6-gamma	391-671
Ps.hal7	77361648	M	F6-gamma	346-626
Ps.hal8	77361657	M	F6-gamma	395-675
Ps.hal9	77362042	M		259-539

Table A.12 (continued)

ID	GI	Gene Neighborhood	Class	Range
Ps.hal10	77362101	M	F6-gamma	426-706
Ps.hal11	77362427	M	F6-gamma	349-629
Ps.put1	26987059	m--M	F6-gamma	261-541
Ps.put2	26987062	m--M	F6-gamma	365-645
Ps.put3	26987300	M	F6-gamma	373-653
Ps.put4	26987322	M	F6-gamma	367-647
Ps.put5	26987963	M	F6-gamma	408-688
Ps.put6	26988105	M	F6-gamma	344-624
Ps.put7	26988221	MW1R1W2A1B1	Alt	260-540
Ps.put8	26988549	M	F6-gamma	366-646
Ps.put9	26988667	M	F6-gamma	436-716
Ps.put10	26988835	M	F6-gamma	241-521
Ps.put11	26988844	M	F6-gamma	275-555
Ps.put12	26988973	M	F6-gamma	363-643
Ps.put13	26988981	M	F6-gamma	241-521
Ps.put14	26989034	M	F6-gamma	212-492
Ps.put15	26989362	M	F6-gamma	270-550
Ps.put16	26989542	M	F6-gamma	270-550
Ps.put17	26989580	M	F6-gamma	264-544
Ps.put18	26990269	M	F6-gamma	434-714
Ps.put19	26991206	M	F6-gamma	241-521
Ps.put20	26991342	M	F6-gamma	359-639
Ps.put21	26991566	M	F6-gamma	397-677
Ps.put22	26991666	Y3Y2W6MA3W5	Tfp	399-679
Ps.put23	26991696	MM	F6-gamma	358-638
Ps.put24	26991697	MM	F6-gamma	367-647
Ps.syr1	28867357	M	F6-beta	1-281
Ps.syr2	28867495	M	F6-gamma	218-498
Ps.syr3	28867696	M	F6-gamma	261-541
Ps.syr4	28868215	M	F6-gamma	277-557
Ps.syr5	28868276	m--m-M	F6-gamma	349-629
Ps.syr6	28868278	m-M--M	F6-gamma	346-626
Ps.syr7	28868281	m-M--M		404-684
Ps.syr8	28868542	M	F6-gamma	269-549
Ps.syr9	28868700	MW2R2W3A2B2	Alt	260-540
Ps.syr10	28868854	M	F6-gamma	241-521
Ps.syr11	28869218	M	F6-gamma	241-521
Ps.syr12	28869455	M	F6-gamma	211-491
Ps.syr13	28869643	M-----w6m	F6-gamma	366-646
Ps.syr14	28869665	M--M----M	F6-gamma	381-661
Ps.syr15	28869668	M--M----M	F6-mix	146-426
Ps.syr16	28869673	M--M----M	F6-gamma	349-629
Ps.syr17	28869704	M	F6-gamma	372-652
Ps.syr18	28869719	M	F6-gamma	279-559
Ps.syr19	28869806	M	F6-gamma	353-633
Ps.syr20	28870175	M	F6-gamma	261-541

Table A.12 (continued)

ID	GI	Gene Neighborhood	Class	Range
Ps.syr21	28870271	M	F6-gamma	362-642
Ps.syr22	28870402	M	F6-gamma	490-770
Ps.syr23	28870443	M	F6-gamma	261-541
Ps.syr24	28870455	M	F6-gamma	262-542
Ps.syr25	28870543	M	F6-gamma	279-559
Ps.syr26	28870737	M--m	F6-gamma	262-542
Ps.syr27	28870740	M--m	F6-gamma	266-546
Ps.syr28	28870835	m----M	F6-gamma	396-676
Ps.syr29	28870840	m----M	F6-gamma	265-545
Ps.syr30	28870854	M	F6-gamma	263-543
Ps.syr31	28871675	M	F6-gamma	261-541
Ps.syr32	28871756	M	F6-gamma	259-539
Ps.syr33	28872050	M	F6-gamma	364-644
Ps.syr34	28872145	Y4Y3W8MA4W7	Tfp	393-673
Ps.syr35	28872271	MM	F6-gamma	358-638
Ps.syr36	28872272	MM	F6-gamma	366-646
Ps.syr37	28872658	MM	F6-gamma	360-640
Ps.syr38	28872659	MM	F6-gamma	394-674
Ps.syr39	28872674	M	F6-gamma	261-541
Ra.eut1	73537488	MW2R1W1A1B1	Alt	259-539
Ra.eut2	73542305	Y5Y4W5MA3	Tfp	452-732
Ra.met1	94309617	Y1Y2W1MA1	Tfp	452-732
Ra.met2	94312901	MW5R3W4A3B2	Alt	326-606
Ra.sol	17545390	Y1Y2W1MA1	Tfp	462-742
Rh.fer	89900168	Y3Y4W6MA3	Tfp	478-758
Sa.deg1	90019730	M	F6-gamma	264-544
Sa.deg2	90019992	M	F6-gamma	267-547
Sa.deg3	90020040	C-M	F6-gamma	219-499
Sa.deg4	90020169	Y2Y1W1MA1	Tfp	408-688
Sa.deg5	90020318	M	F6-gamma	251-531
Sa.deg6	90020755	M	F6-gamma	259-539
Sa.deg7	90021961	M	F6-gamma	266-546
Sa.deg8	90023065	M	F6-gamma	136-416
Sa.deg9	90023271	Y7Y6W6MR4A4B4W5	Tfp	389-669
Sa.rub1	83814438	M	Sa.rub	1-281
Sa.rub2	83814592	M	Sa.rub	1-281
Sa.rub3	83814728	MM	Sa.rub	1-281
Sa.rub4	83814750	M	Sa.rub	1-281
Sa.rub5	83815042	MM	Sa.rub	1-281
Sa.rub6	83815551	M	Sa.rub	1-281
Sa.rub7	83815947	WR2Y-AB1-M-M-M	Sa.rub	597-877
Sa.rub8	83816322	WR2Y-AB1-M-M-M	Sa.rub	350-630
Sa.rub9	83816813	M	Sa.rub	157-437
Sh.den1	91791468	M	F6-gamma	359-639
Sh.den2	91792074	M	F6-gamma	260-540
Sh.den3	91792221	M	F6-gamma	201-481

Table A.12 (continued)

ID	GI	Gene Neighborhood	Class	Range
Sh.den4	91792271	M	F6-gamma	425-705
Sh.den5	91792451	M	F6-gamma	340-620
Sh.den6	91792956	MM	F6-gamma	258-538
Sh.den7	91792957	MM	F6-gamma	220-500
Sh.den8	91793318	M	F6-gamma	254-534
Sh.den9	91793591	M	F6-gamma	368-648
Sh.den10	91793787	M	F6-gamma	111-391
Sh.den11	91793989	M	F6-gamma	395-675
Sh.den12	91794242	M	F6-gamma	393-673
Sh.den13	91794305	M	F6-gamma	217-497
Sh.den14	91794360	M	F6-gamma	259-539
Sh.den15	91794416	M	F6-gamma	198-478
Sh.den16	91794449	M		372-652
Sh.one1	24372094	M	F6-beta	262-542
Sh.one2	24372177	M	F6-gamma	1-281
Sh.one3	24372574	M	F6-gamma	274-554
Sh.one4	24372641	M	F6-gamma	350-630
Sh.one5	24372859	M	F6-gamma	426-706
Sh.one6	24372963	M	F6-gamma	1-281
Sh.one7	24373012	M	F6-gamma	259-539
Sh.one8	24373643	M	F6-gamma	225-505
Sh.one9	24373793	x1M	F6-gamma	265-545
Sh.one10	24374574	M	F6-gamma	299-579
Sh.one11	24374793	M	F6-gamma	387-667
Sh.one12	24374914	M	F6-gamma	235-515
Sh.one13	24375141	M	F6-gamma	393-673
Sh.one14	24375328	M	F6-gamma	340-620
Sh.one15	24375378	M	F6-gamma	103-383
Sh.one16	24375540	M	F6-gamma	354-634
Sh.one17	24375932	M	F6-gamma	362-642
Sh.one18	24375944	M	F6-gamma	371-651
Sh.one19	24376031	M	F6-gamma	343-623
Sh.one20	24376324	M	F6-gamma	266-546
Sh.one21	50261353	M	F6-gamma	261-541
Si.mel	16263300	R2W4MA2B2	Alt	276-556
Sy.elo1	56750540	W1MA1	Tfp	551-831
Sy.elo2	56750691	Y3Y2W3MA2W2	Tfp	1122-1402
Synco1	86606796	Y3Y2W1---MA1	Tfp	69-349
Synco2	86607345	Y4Y5Y6W2MA2	Tfp	533-813
Syncy1	16329620	Y1Y2W1M	Tfp	586-866
Syncy2	16329791	Y4Y3W2MA1	Tfp	672-952
Syncy3	16331988	Y6Y5W4MMA3W3	Tfp	719-999
Tb.den	74318567	Y4Y3W2MA2	Tfp	426-706
Th.elo1	22297891	Y1Y2W1MA1	Tfp	767-1047
Th.elo2	22298112	Y4Y3W3MA2W2	Tfp	656-936
Th.elo3	22298566	Y6Y5W4MA3	Tfp	237-517

Table A.12 (continued)

ID	GI	Gene Neighborhood	Class	Range
Tm.cru1	78484915	M---M	F6-gamma	394-674
Tm.cru2	78486338	M----m	F6-gamma	123-403
Tm.den1	78777186	mM----m-mz1-b1d-r1a1-mw1y1	F3	1-281
Tm.den2	78778126	M	F3	402-682
Vi.cho1	15600779	M	F6-gamma	273-553
Vi.cho2	15600802	M		241-521
Vi.cho3	15600839	M	F6-gamma	267-547
Vi.cho4	15600946	M	F6-gamma	364-644
Vi.cho5	15600989	M	F6-gamma	268-548
Vi.cho6	15601036	M	F6-gamma	362-642
Vi.cho7	15601416	m----M	F6-gamma	255-535
Vi.cho8	15601421	m----M	F6-gamma	304-584
Vi.cho9	15601528	M	F6-gamma	193-473
Vi.cho10	15601660	M	F6-gamma	298-578
Vi.cho11	15601677	M	F6-gamma	372-652
Vi.cho12	15601727	M----M	F6-gamma	281-561
Vi.cho13	15601732	M----M	F6-gamma	14-294
Vi.cho14	15601741	M	F6-gamma	237-517
Vi.cho15	15601786	M	F6-gamma	386-666
Vi.cho16	15601807	M	F6-gamma	490-770
Vi.cho17	15601820	M	F6-gamma	359-639
Vi.cho18	15640246	M	F6-gamma	284-564
Vi.cho19	15640311	M	F6-gamma	276-556
Vi.cho20	15640476	M	F6-gamma	259-539
Vi.cho21	15640536	M-M	F6-gamma	246-526
Vi.cho22	15640538	M-M	F6-gamma	346-626
Vi.cho23	15640842	M	F6-gamma	336-616
Vi.cho24	15640857	M	F6-gamma	346-626
Vi.cho25	15641261	M	F6-gamma	256-536
Vi.cho26	15641302	M	F6-gamma	351-631
Vi.cho27	15641311	M	F6-gamma	253-533
Vi.cho28	15641325	M	F6-gamma	262-542
Vi.cho29	15641416	m-a2y3r2-b1w3m-MM	F6-gamma	314-594
Vi.cho30	15641424	M	F6-gamma	297-577
Vi.cho31	15641543	M	F6-gamma	249-529
Vi.cho32	15641648	M	F6-gamma	316-596
Vi.cho33	15641861	M	F6-gamma	425-705
Vi.cho34	15641870	M		346-626
Vi.cho35	15641900	M	F6-gamma	392-672
Vi.cho36	15641969	M	F6-gamma	404-684
Vi.cho37	15642160	M	F6-gamma	343-623
Vi.cho38	15642435	M	F6-gamma	153-433
Vi.fis1	59711305	M	F6-gamma	382-662
Vi.fis2	59711384	M	F6-gamma	344-624
Vi.fis3	59711434	M	F6-gamma	271-551
Vi.fis4	59711594	M	F6-gamma	369-649

Table A.12 (continued)

ID	GI	Gene Neighborhood	Class	Range
Vi.fis5	59711698	MM	F6-gamma	244-524
Vi.fis6	59711699	MM	F6-gamma	387-667
Vi.fis7	59711724	M	F6-gamma	321-601
Vi.fis8	59711740	M----M	F6-gamma	389-669
Vi.fis9	59711745	M----M	F6-gamma	339-619
Vi.fis10	59711976	M	F6-gamma	397-677
Vi.fis11	59712110	MM	F6-gamma	187-467
Vi.fis12	59712111	MM	F6-gamma	372-652
Vi.fis13	59712225	M	F6-gamma	266-546
Vi.fis14	59712259	M	F6-gamma	345-625
Vi.fis15	59712385	M	F6-gamma	259-539
Vi.fis16	59712396	M	F6-gamma	416-696
Vi.fis17	59712649	M	F6-gamma	262-542
Vi.fis18	59712768	M	F6-gamma	1-281
Vi.fis19	59712843	M	F6-gamma	253-533
Vi.fis20	59713275	M	F6-gamma	237-517
Vi.fis21	59713290	M	F6-gamma	326-606
Vi.fis22	59713352	MM	F6-gamma	263-543
Vi.fis23	59713353	MM	F6-gamma	261-541
Vi.fis24	59713429	M	F6-gamma	341-621
Vi.fis25	59713483	M-M	F6-gamma	381-661
Vi.fis26	59713485	M-M	F6-gamma	372-652
Vi.fis27	59713508	M	F6-gamma	281-561
Vi.fis28	59713572	M	F6-gamma	341-621
Vi.fis29	59713630	MM	F6-gamma	262-542
Vi.fis30	59713631	MM	F6-gamma	262-542
Vi.fis31	59713664	M	F6-gamma	267-547
Vi.fis32	59713710	MM	F6-gamma	250-530
Vi.fis33	59713711	MM	F6-gamma	267-547
Vi.fis34	59713860	M	F6-gamma	357-637
Vi.fis35	59714042	M-----M	F6-gamma	343-623
Vi.fis36	59714048	M-----M	F6-gamma	351-631
Vi.fis37	59714074	M	F6-gamma	1-281
Vi.fis38	59714252	mmM-M	F6-gamma	249-529
Vi.fis39	59714254	mmM-M	F6-gamma	1-281
Vi.fis40	59714255	m-mMM	F6-gamma	244-524
Vi.fis41	59714256	m-mMM	F6-gamma	252-532
Vi.fis42	59714267	M	F6-gamma	343-623
Vi.par1	28896957	M	F6-gamma	277-557
Vi.par2	28897196	M	F6-gamma	162-442
Vi.par3	28897737	M	F6-gamma	383-663
Vi.par4	28897862	M	F6-gamma	265-545
Vi.par5	28897959	M		258-538
Vi.par6	28898260	C1---M	F6-gamma	247-527
Vi.par7	28898402	M	F6-gamma	187-467
Vi.par8	28898666	M	F6-gamma	253-533

Table A.12 (continued)

ID	GI	Gene Neighborhood	Class	Range
Vi.par9	28898678	M	F6-gamma	347-627
Vi.par10	28898755	M	F6-gamma	417-697
Vi.par11	28898933	M	F6-gamma	1-281
Vi.par12	28899403	M	F6-gamma	259-539
Vi.par13	28899601	M	F6-gamma	346-626
Vi.par14	28899879	M	F6-gamma	273-553
Vi.par15	28900054	M	F6-gamma	384-664
Vi.par16	28900346	M	F6-gamma	262-542
Vi.par17	28900366	M	F6-gamma	343-623
Vi.par18	28900417	M	F6-gamma	399-679
Vi.par19	28900451	M	F6-gamma	1-281
Vi.par20	28900697	M	F6-gamma	195-475
Vi.par21	28900855	M	F6-gamma	262-542
Vi.par22	28901037	M	F6-gamma	239-519
Vi.par23	28901044	M	F6-gamma	306-586
Vi.par24	28901304	M	F6-gamma	358-638
Vi.par25	28901317	M	F6-gamma	267-547
Vi.par26	28901347	M	F6-gamma	262-542
Vi.par27	28901506	M	F6-gamma	347-627
Vi.vul1	27363527	M	F6-gamma	264-544
Vi.vul2	27363668	M	F6-gamma	346-626
Vi.vul3	27364074	M	F6-gamma	138-418
Vi.vul4	27364275	M	F6-gamma	275-555
Vi.vul5	27364632	M	F6-gamma	266-546
Vi.vul6	27364882	M	F6-gamma	1-281
Vi.vul7	27365130	M	F6-gamma	247-527
Vi.vul8	27365157	M	F6-gamma	255-535
Vi.vul9	27365391	M	F6-gamma	1-281
Vi.vul10	27365400	M	F6-gamma	375-655
Vi.vul11	27365429	M		346-626
Vi.vul12	27365442	M	F6-gamma	1-281
Vi.vul13	27365528	M		266-546
Vi.vul14	27365568	M	F6-gamma	347-627
Vi.vul15	27365579	M	F6-gamma	253-533
Vi.vul16	27365673	M	F6-gamma	233-513
Vi.vul17	27365863	C1---M	F6-gamma	394-674
Vi.vul18	27365879	M	F6-gamma	340-620
Vi.vul19	27366002	M	F6-gamma	187-467
Vi.vul20	27366013	M	F6-gamma	239-519
Vi.vul21	27366036	M	F6-gamma	297-577
Vi.vul22	27366148	M		255-535
Vi.vul23	27366407	M	F6-gamma	1-281
Vi.vul24	27366502	M	F6-gamma	354-634
Vi.vul25	27366531	M---v3	F6-gamma	238-518
Vi.vul26	27366677	M	F6-gamma	181-461
Vi.vul27	27366778	M	F6-gamma	490-770

Table A.12 (continued)

ID	GI	Gene Neighborhood	Class	Range
Vi.vul28	27366798	M	F6-gamma	258-538
Vi.vul29	27366812	M----R2B2	F6-gamma	1-281
Vi.vul30	27366840	MM	F6-gamma	341-621
Vi.vul31	27366841	MM	F6-gamma	347-627
Vi.vul32	27366911	M	F6-gamma	262-542
Vi.vul33	27366961	M	F6-gamma	279-559
Vi.vul34	27366987	M	F6-gamma	239-519
Vi.vul35	27367232	M	F6-gamma	216-496
Vi.vul36	27367374	M	F6-gamma	301-581
Vi.vul37	27367433	M	F6-gamma	381-661
Vi.vul38	27367488	M	F6-gamma	368-648
Vi.vul39	27367575	M	F6-gamma	250-530
Vi.vul40	27367630	M	F6-gamma	366-646
Vi.vul41	27367639	M	F6-gamma	260-540
Vi.vul42	27367871	M	F6-gamma	1-281
Vi.vul43	27367899	M-M	F6-gamma	183-463
Vi.vul44	27367901	M-M	F6-gamma	267-547
Vi.vul45	27367913	M	F6-gamma	357-637
Vi.vul46	27367965	MM---V4	F6-gamma	342-622
Vi.vul47	27367966	MM---V4	F6-gamma	343-623
Vi.vul48	27367986	M	F6-gamma	384-664
Wo.suc1	34557060	M	F3	261-541
Wo.suc2	34557162	M	F3	236-516
Wo.suc3	34557253	M----M		265-545
Wo.suc4	34557389	M	F3	387-667
Wo.suc5	34557756	M	F3	348-628
Wo.suc6	34558217	M	F3	1-281
Wo.suc7	34558397	M	F3	250-530
Xa.axo	21243826	Y6Y5W4MA3B4W3	Tfp	396-676
Xa.cam	21232353	Y6Y5W5MA4B5W4	Tfp	396-676
Xa.ory	58581372	Y6Y2W2MA2A2B2W3	Tfp	396-676
Xy.fas	15838547	Y2W2MABW1	Tfp	411-691

Table A.13 MCP data. ID, GI, and Gene Neighborhood are explained in Table A.2. MCP length class was determined by HMM analysis.

ID	GI	Gene Neighborhood	Class
Ac.bac	94968554	A1W1MY1B1R1	38H
Ac.bac	94968699	M	38H
Ac.bac	94968801	MW2R2A2B2Y2X1	
Ag.tum	15887724	M	36H
Ag.tum	15887736	M	36H
Ag.tum	15887862	M-Y1ARBY2D-----M	36H
Ag.tum	15887875	M-Y1ARBY2D-----M	36H
Ag.tum	15887987	M	36H
Ag.tum	15888080	M	36H
Ag.tum	15888213	M	36H
Ag.tum	15888368	M	36H
Ag.tum	15889206	M	36H
Ag.tum	15889452	M	36H
Ag.tum	15889499	M	36H
Ag.tum	15889634	M	36H
Ag.tum	15889874	W2M	36H
Ag.tum	15890257	M	36H
Ag.tum	15891223	M	36H
Ag.tum	15891570	M	36H
Ag.tum	15891604	M	36H
Ag.tum	15891832	M	36H
Ag.tum	16119670	M	36H
Ag.tum	16119950	M	36H
An.var	75906286	Y1Y2W1MA1	40H
An.var	75906729	Y4Y3W2MA2	40H
An.var	75909982	Y6Y7W3MA3	40H
An.deh	86156460	M	34H
An.deh	86156488	M	64H
An.deh	86156754	M	36H
An.deh	86157040	b1r1w1a1d1-y2--y3W2W3MA2B2R2	40H
An.deh	86157103	M	
An.deh	86157586	M	34H
An.deh	86157620	R3W4MA3B3	
An.deh	86157800	b5r5w6ma5y5--W5MB4d2r4a4	36H
An.deh	86157806	A4R4D2b4mw5--Y5A5MW6R5B5	34H
An.deh	86158609	MW7	34H
An.deh	86159505	M	34H
An.deh	86159583	M	36H
An.deh	86159820	M	
An.deh	86159931	M	34H
An.deh	86160122	M	34H
An.deh	86160136	M	34H
An.deh	86160313	M	34H
An.deh	86160734	MW11Y8A7C	44H
Ar.ful	11498639	MW-YBACDR--M	44H

Table A.13 (continued)

ID	GI	Gene Neighborhood	Class
Ar.ful	11498650	MW-YBACDR--M	44H
Azoar	56476384	Y1Y2WMA	40H
Azoar	56478990	M	40H
Ba.ant	49183383	mM	44H
Ba.ant	49183384	mM	44H
Ba.ant	49183393	M	44H
Ba.ant	49183552	M	44H
Ba.ant	49183570	M	44H
Ba.ant	49183675	M	
Ba.ant	49184066	M	44H
Ba.ant	49184879	M	44H
Ba.ant	49186061	M	44H
Ba.ant	49187871	M	44H
Ba.ant	49187893	M	44H
Ba.ant	49187927	M	
Ba.ant	49188261	M	
Ba.cer	30018612	M	44H
Ba.cer	30018630	M	44H
Ba.cer	30018745	M	44H
Ba.cer	30018762	M	44h
Ba.cer	30018860	M	
Ba.cer	30019279	M	44H
Ba.cer	30020143	M	44H
Ba.cer	30021489	M	44H
Ba.cer	30021622	M	44H
Ba.cer	30023049	M	44H
Ba.cer	30023073	M	44H
Ba.cer	30023102	M	44H
Ba.cer	30023454	M	
Ba.cla	56962237	M	44H
Ba.cla	56963771	M	44H
Ba.cla	56965241	M	
Ba.cla	56965295	M	44H
Ba.cla	56965624	M	44H
Ba.cla	56965723	M	44H
Ba.cla	56965807	M	44H
Ba.hal	15613068	M	
Ba.hal	15613128	M	44H
Ba.hal	15613439	M	44H
Ba.hal	15613569	M	
Ba.hal	15614072	M	
Ba.hal	15614390	M	44H
Ba.hal	15614838	M	44H
Ba.hal	15616330	M	44H
Ba.hal	15616425	M	44H

Table A.13 (continued)

ID	GI	Gene Neighborhood	Class
Ba.hal	15616462	M--M	44H
Ba.hal	15616465	M--M	44H
Ba.hal	15616477	M	44H
Ba.lic	52078713	M	44H
Ba.lic	52078865	M	44H
Ba.lic	52078907	M	44H
Ba.lic	52079190	M	
Ba.lic	52079522	M	
Ba.lic	52079992	M	44H
Ba.lic	52080553	M	44H
Ba.lic	52081612	MMM	44H
Ba.lic	52081613	MMM	44H
Ba.lic	52081614	MMM	44H
Ba.lic	52082555	M	44H
Ba.sub	16077413	M	44H
Ba.sub	16077803	M	
Ba.sub	16078102	M	
Ba.sub	16078459	M-----V	44H
Ba.sub	16078921	M	44H
Ba.sub	16080175	MMMM	44H
Ba.sub	16080176	MMMM	44H
Ba.sub	16080177	MMMM	44H
Ba.sub	16080178	MMMM	44H
Ba.sub	16080422	M	44H
Ba.thu	49476793	mM	44H
Ba.thu	49478863	MM	44H
Ba.thu	49478864	MM	44H
Ba.thu	49479039	M	44H
Ba.thu	49480036	mM	44H
Ba.thu	49480152	M	44H
Ba.thu	49480166	M	44H
Ba.thu	49480248	M	
Ba.thu	49481071	M	44H
Ba.thu	49481103	M	44H
Ba.thu	49481221	M	44H
Ba.thu	49481265	M	
Bd.bac	42521760	M	36H
Bd.bac	42521886	M	36H
Bd.bac	42522250	M	36H
Bd.bac	42522500	M	36H
Bd.bac	42522635	M	36H
Bd.bac	42522674	M	36H
Bd.bac	42522860	M	36H
Bd.bac	42522982	MM	36H
Bd.bac	42522983	MM	36H
Bd.bac	42523352	M	

Table A.13 (continued)

ID	GI	Gene Neighborhood	Class
Bd.bac	42523931	Mm	36H
Bd.bac	42523932	Mm	36H
Bd.bac	42524018	M	36H
Bd.bac	42524042	M	36H
Bd.bac	42524240	MR1D2W2	
Bd.bac	42524481	M	
Bd.bac	42524560	M	36H
Bd.bac	42524574	M	36H
Bd.bac	42524632	M	36H
Bd.bac	42524709	M	36H
Bo.bro	33600194	M	36H
Bo.bro	33600441	M	36H
Bo.bro	33601527	Y1AWMRBY2Z	36H
Bo.bro	33601549	M-MM	36H
Bo.bro	33601551	M-MM	36H
Bo.bro	33601552	M-MM	36H
Bo.bro	33601934	M	36H
Bo.bro	33602091	M	36H
Bo.par	33595672	M	36H
Bo.par	33596128	Y1AWMRBY2Z	36H
Bo.par	33596147	M	36H
Bo.par	33596249	M	36H
Bo.par	33596677	M	36H
Bo.par	33597536	M	36H
Bo.per	33592175	AWMBYZ	36H
Bo.per	33592487	MM	36H
Bo.per	33592488	MM	36H
Bo.per	33592680	M	36H
Bo.per	33593267	M	36H
Bo.bur	11496688	M	
Bo.bur	15594923	M	
Bo.bur	15594941	mM	48H
Bo.bur	15594942	mM	
Bo.bur	15595025	MM	34H
Bo.bur	15595026	MM	34H
Bo.gar	51598830	M	
Bo.gar	51598849	mM	48H
Bo.gar	51598850	mM	
Bo.gar	51598935	MM	34H
Bo.gar	51598936	MM	34H
Br.jap	27375207	M	38H
Br.jap	27375437	MM	38H
Br.jap	27375438	MM	38H
Br.jap	27375494	M	38H
Br.jap	27375687	M	38H
Br.jap	27376034	M-----M	24H

Table A.13 (continued)

ID	GI	Gene Neighborhood	Class
Br.jap	27376040	M-----M	38H
Br.jap	27376314	M	38H
Br.jap	27376581	MM	38H
Br.jap	27376582	MM	38H
Br.jap	27376643	MM	
Br.jap	27376644	MM	
Br.jap	27377456	Y3A3W3MW4MR3B2	34H
Br.jap	27377458	Y3A3W3MW4MR3B2	34H
Br.jap	27377617	MM	38H
Br.jap	27377618	MM	38H
Br.jap	27377657	MMM	38H
Br.jap	27377658	MMM	38H
Br.jap	27377659	MMM	38H
Br.jap	27378042	MMM	38H
Br.jap	27378043	MMM	38H
Br.jap	27378044	MMM	38H
Br.jap	27378087	M	38H
Br.jap	27378104	M	38H
Br.jap	27378240	M	38H
Br.jap	27379302	M----m	
Br.jap	27379307	M----m	38H
Br.jap	27379437	MM	38H
Br.jap	27379438	MM	38H
Br.jap	27379487	M	
Br.jap	27380420	M	38H
Br.jap	27382057	M	38H
Br.jap	27382173	M	38H
Br.jap	27382284	M	38H
Br.jap	27382406	M	38H
Br.jap	27383065	M	38H
Bu.cen	107022518	M	36H
Bu.cen	107022684	M	36H
Bu.cen	107023768	M	36H
Bu.cen	107024392	Y2A1W1MR1DB1Y1Z	36H
Bu.cen	107025360	MM	36H
Bu.cen	107025361	MM	36H
Bu.cen	107025593	M	
Bu.cen	107025900	M	36H
Bu.cen	107026057	M	36H
Bu.cen	107026277	M	36H
Bu.cen	107026343	M	36H
Bu.cen	107026421	M	36H
Bu.cen	107026645	M--Y3	
Bu.cen	107026920	MW3R2-W2A2B3	40H
Bu.cen	107026934	M	36H
Bu.cen	107026970	M	36H

Table A.13 (continued)

ID	GI	Gene Neighborhood	Class
Bu.cen	107027695	M	36H
Bu.cen	107028433	M	36H
Bu.cen	107029185	M	36H
Bu.cen	107029202	mM	36H
Bu.cen	107029203	mM	36H
Bu.mal	53716046	M	36H
Bu.mal	53716190	M	36H
Bu.mal	53716498	MW2R1W1A1B2	40H
Bu.mal	53716753	M	36H
Bu.mal	53716774	M	36H
Bu.mal	53716899	M	36H
Bu.mal	53717276	M	36H
Bu.mal	53723395	MW3	36H
Bu.mal	53723482	M	36H
Bu.mal	53724320	Y4A2W4MR2DB1-Y3Z	36H
Bu.mal	53724360	M	36H
Bu.mal	53724505	M	36H
Bu.mal	53724713	M	36H
Bu.mal	53724892	M	36H
Bu.mal	53725324	M	36H
Bu.mal	53725441	M	36H
Bu.mal	53725699	M	36H
Bu.pse	53717964	M	36H
Bu.pse	53718105	M	36H
Bu.pse	53718349	M	36H
Bu.pse	53719241	M	36H
Bu.pse	53719346	MW1	36H
Bu.pse	53719442	M	36H
Bu.pse	53719488	M	36H
Bu.pse	53719588	M	36H
Bu.pse	53719974	M	36H
Bu.pse	53720914	Y2A1W2MR1DB1Y1Z	36H
Bu.pse	53720948	M	36H
Bu.pse	53721239	M	36H
Bu.pse	53721250	M	36H
Bu.pse	53721311	M	36H
Bu.pse	53721497	M	36H
Bu.pse	53721759	M	36H
Bu.pse	53721957	M	36H
Bu.pse	53722241	M	36H
Bu.pse	53722895	MW4R2W3A2B2	40H
Bu.pse	53722948	M	36H
Bu.pse	53722962	M	36H
Burkh	78059760	M	36H
Burkh	78060620	M	36H
Burkh	78060711	M	36H

Table A.13 (continued)

ID	GI	Gene Neighborhood	Class
Burkh	78061010	M	36H
Burkh	78061119	MM	36H
Burkh	78061120	MM	36H
Burkh	78061731	M	36H
Burkh	78062008	M	36H
Burkh	78062373	M	36H
Burkh	78062492	M	36H
Burkh	78062827	M--Y2	
Burkh	78063155	MW2R1W1A1B2	40H
Burkh	78063204	M	36H
Burkh	78064837	Y3A2W3MR2DB3Y4Z	36H
Burkh	78065556	M	36H
Burkh	78066060	M	36H
Burkh	78066220	M	36H
Burkh	78067633	M	36H
Bu.tha	83716119	M	36H
Bu.tha	83716541	MW1R2W3A2B3	40H
Bu.tha	83716741	M	36H
Bu.tha	83716986	M	36H
Bu.tha	83717148	Y3A1W2MR1D1B2B1	36H
Bu.tha	83717565	M	36H
Bu.tha	83717668	M	36H
Bu.tha	83717947	M	36H
Bu.tha	83717972	M	36H
Bu.tha	83718188	M	36H
Bu.tha	83718421	M	36H
Bu.tha	83718725	M	36H
Bu.tha	83718995	M	36H
Bu.tha	83719305	M	36H
Bu.tha	83719352	M	36H
Bu.tha	83720415	M	36H
Bu.tha	83720865	Y4A3W4MR3D2B4Y5Z	36H
Bu.tha	83720882	M	36H
Bu.tha	83721176	MW5	36H
Bu.tha	83721567	M	36H
Bu.xen	91777129	M--Y1	
Bu.xen	91777154	M	36H
Bu.xen	91777185	M	36H
Bu.xen	91777198	M	36H
Bu.xen	91777353	M	36H
Bu.xen	91777453	M	36H
Bu.xen	91777507	M	36H
Bu.xen	91777613	M	36H
Bu.xen	91777639	M	36H
Bu.xen	91777891	M	36H
Bu.xen	91778971	M	

Table A.13 (continued)

ID	GI	Gene Neighborhood	Class
Bu.xen	91779325	M	36H
Bu.xen	91779951	M	36H
Bu.xen	91780080	M	
Bu.xen	91780653	M	36H
Bu.xen	91780869	M	36H
Bu.xen	91780931	M	36H
Bu.xen	91781077	M	36H
Bu.xen	91781147	M	36H
Bu.xen	91781544	M	36H
Bu.xen	91781687	M	36H
Bu.xen	91781765	M	36H
Bu.xen	91782104	M	36H
Bu.xen	91783410	M	36H
Bu.xen	91783789	M	36H
Bu.xen	91783810	M	36H
Bu.xen	91783857	M	
Bu.xen	91783988	M	36H
Bu.xen	91784395	MM	36H
Bu.xen	91784396	MM	36H
Bu.xen	91784458	M	36H
Bu.xen	91785656	Y4AWMR4DB4Y3Z	36H
Ca.jej	15791418	M	
Ca.jej	15791532	M	28H
Ca.jej	15791617	M	
Ca.jej	15791633	M	28H
Ca.jej	15791812	M	
Ca.jej	15792280	M	
Ca.jej	15792435	M	24H
Ca.jej	15792514	M	
Ca.jej	15792820	M	28H
Ca.jej	15792869	M	28H
Ca.hyd	78042826	DR2CY1----M-MY3-R1MW1B2Y2A2	
Ca.hyd	78042845	MM	
Ca.hyd	78042880	M	
Ca.hyd	78043239	MM	
Ca.hyd	78043413	M	
Ca.hyd	78043540	M	44H
Ca.hyd	78043972	MW2A1B1	
Ca.hyd	78044371	M	
Ca.hyd	78044569	DR2CY1----M-MY3-R1MW1B2Y2A2	
Ca.hyd	78044580	M	
Ca.hyd	78044786	DR2CY1----M-MY3-R1MW1B2Y2A2	44H
Ca.cre	16124321	M	36H
Ca.cre	16124598	M	36H
Ca.cre	16124683	M-M-Y1A1W1R1B1Y2D	36H
Ca.cre	16124685	M-M-Y1A1W1R1B1Y2D	36H

Table A.13 (continued)

ID	GI	Gene Neighborhood	Class
Ca.cre	16124759	M	36H
Ca.cre	16124843	r2b2y4w2a2m-Y3ZM	36H
Ca.cre	16124847	mzy3-MA2W2Y4B2R2	38H
Ca.cre	16125648	M	36H
Ca.cre	16125901	M	36H
Ca.cre	16126520	M	36H
Ca.cre	16126556	M	36H
Ca.cre	16126924	M	36H
Ca.cre	16127042	M	
Ca.cre	16127074	m----M	36H
Ca.cre	16127079	m----M	36H
Ca.cre	16127375	M	36H
Ca.cre	16127579	M	36H
Ca.cre	16127588	M	36H
Ch.vio	34495486	M	40H
Ch.vio	34495550	M	40H
Ch.vio	34495712	M	40H
Ch.vio	34495720	M	40H
Ch.vio	34495850	M	
Ch.vio	34496082	M	40H
Ch.vio	34496168	M	40H
Ch.vio	34496328	M	40H
Ch.vio	34496354	M	40H
Ch.vio	34496466	Y1-A1MW1MD1B1	34H
Ch.vio	34496468	Y1-A1MW1MD1B1	34H
Ch.vio	34496535	MM	40H
Ch.vio	34496536	MM	40H
Ch.vio	34496783	M	40H
Ch.vio	34496813	M	40H
Ch.vio	34496872	M	40H
Ch.vio	34496909	M	24H
Ch.vio	34496979	M	40H
Ch.vio	34497107	M	34H
Ch.vio	34497132	M	34H
Ch.vio	34497153	M	24H
Ch.vio	34497171	M	40H
Ch.vio	34497305	M	40H
Ch.vio	34497367	M---M	
Ch.vio	34497371	M---M	
Ch.vio	34497918	M	40H
Ch.vio	34497964	MY2-A2MW2R2B3	34H
Ch.vio	34497968	MY2-A2MW2R2B3	
Ch.vio	34498101	M	
Ch.vio	34498307	M	40H
Ch.vio	34498314	M	40H
Ch.vio	34498362	M	40H

Table A.13 (continued)

ID	GI	Gene Neighborhood	Class
Ch.vio	34498740	M	40H
Ch.vio	34498775	MM	40H
Ch.vio	34498776	MM	34H
Ch.vio	34498893	a4zy5v3v2--Y4A3W3--MR3B5D2	36H
Ch.vio	34499104	M	40H
Ch.vio	34499321	M	40H
Ch.vio	34499610	MM	40H
Ch.vio	34499611	MM	40H
Ch.vio	34499699	M	40H
Ch.vio	34499752	M	40H
Ch.sal	92112169	M	36H
Ch.sal	92112261	M	36H
Ch.sal	92112311	M	36H
Ch.sal	92113033	MM	24H
Ch.sal	92113034	MM	24H
Ch.sal	92113226	M	36H
Ch.sal	92113548	M	36H
Ch.sal	92113598	M	36H
Ch.sal	92113849	M	36H
Ch.sal	92114111	M	36H
Ch.sal	92114141	AWMRBMYZ	36H
Ch.sal	92114144	AWMRBMYZ	36H
Ch.sal	92114896	M	36H
Cl.ace	15004752	M	
Cl.ace	15893416	Y1A1W1MR1Y2	36H
Cl.ace	15893542	M	44H
Cl.ace	15893672	M	
Cl.ace	15893723	Mm	
Cl.ace	15893724	Mm	44H
Cl.ace	15893734	M	44H
Cl.ace	15893832	MM	
Cl.ace	15893833	MM	44H
Cl.ace	15894028	M	
Cl.ace	15894092	M	44H
Cl.ace	15894102	M	44H
Cl.ace	15894196	M	44H
Cl.ace	15894631	M	44H
Cl.ace	15894666	M	44H
Cl.ace	15894802	M	
Cl.ace	15894857	M	44H
Cl.ace	15894878	MM	
Cl.ace	15894879	MM	
Cl.ace	15895686	M	44H
Cl.ace	15895748	M	
Cl.ace	15895809	M	44H
Cl.ace	15895875	M	44H

Table A.13 (continued)

ID	GI	Gene Neighborhood	Class
Cl.ace	15896002	Mm	
Cl.ace	15896003	Mm	44H
Cl.ace	15896019	M	
Cl.ace	15896029	M	44H
Cl.ace	15896058	M	44H
Cl.ace	15896076	M	44H
Cl.ace	15896488	M	44H
Cl.ace	15896595	M	44H
Cl.ace	15896629	M	
Cl.ace	15896638	M	44H
Cl.ace	15896714	M	
Cl.ace	15896747	MM	44H
Cl.ace	15896748	MM	
Cl.ace	15896781	M	
Cl.ace	15896920	M	44H
Cl.tet	28209944	M	44H
Cl.tet	28210059	M	
Cl.tet	28210223	M	44H
Cl.tet	28210425	M	
Cl.tet	28210574	M	44H
Cl.tet	28210600	M	44H
Cl.tet	28210635	M	44H
Cl.tet	28210653	M	44H
Cl.tet	28210662	M	44H
Cl.tet	28210848	M	
Cl.tet	28210899	M	44H
Cl.tet	28210992	M	44H
Cl.tet	28211163	M	
Cl.tet	28211181	M--M	44H
Cl.tet	28211184	M--M	44H
Cl.tet	28211206	M	44H
Cl.tet	28211474	M--M	44H
Cl.tet	28211477	M--M	44H
Cl.tet	28211494	M	44H
Cl.tet	28211515	M	44H
Co.psy	71277838	M	40H
Co.psy	71278287	M	40H
Co.psy	71278459	M	40H
Co.psy	71278766	M	40H
Co.psy	71278893	M	40H
Co.psy	71279028	M---M	40H
Co.psy	71279083	M--M	40H
Co.psy	71279453	M	40H
Co.psy	71279484	M	40H
Co.psy	71279686	M	40H
Co.psy	71279880	M	40H

Table A.13 (continued)

ID	GI	Gene Neighborhood	Class
Co.psy	71280322	M	40H
Co.psy	71280537	M	40H
Co.psy	71280713	M	40H
Co.psy	71280779	M---M	40H
Co.psy	71280909	M	
Co.psy	71280933	M	40H
Co.psy	71281197	M	40H
Co.psy	71281232	M--M	40H
Co.psy	71281444	M	40H
Co.psy	71281933	M	40H
Co.psy	71282390	M	40H
Co.psy	71282609	M	40H
De.aro	71906201	M	40H
De.aro	71906313	M	40H
De.aro	71906363	MY1A1W1MMR1D1B1-V1V2Y2ZA2	
De.aro	71906367	MY1A1W1MMR1D1B1-V1V2Y2ZA2	36H
De.aro	71906368	MY1A1W1MMR1D1B1-V1V2Y2ZA2	36H
De.aro	71906476	M	40H
De.aro	71906565	M	40H
De.aro	71906779	Y3M-A3MW3D2R2D3B2	
De.aro	71906782	Y3M-A3MW3D2R2D3B2	34H
De.aro	71906995	M	40H
De.aro	71907531	M	40H
De.aro	71908114	M	
De.aro	71908341	M	40H
De.aro	71908377	M	24H
De.aro	71908447	M	40H
De.aro	71908456	M	40H
De.aro	71908496	M	40H
De.aro	71908546	M	40H
De.aro	71908710	x----M	40H
De.aro	71908725	M	40H
De.aro	71908738	M	
De.aro	71909191	M	40H
De.aro	71909422	M	40H
De.aro	71909507	Y7Y6W4MA4	40H
De.aro	71909618	M	
De.rad	15808011	YWMMMA	40H
De.rad	15808012	YWMMMA	40H
De.rad	15808013	YWMMMA	40H
De.haf	89892885	MM	44H
De.haf	89892886	MM	44H
De.haf	89892966	M	44H
De.haf	89893096	M	44H
De.haf	89893207	M	
De.haf	89893313	M	44H

Table A.13 (continued)

ID	GI	Gene Neighborhood	Class
De.haf	89893325	M--M	44H
De.haf	89893328	M--M	
De.haf	89893403	M	44H
De.haf	89893637	M	
De.haf	89893670	M	44H
De.haf	89893745	M	44H
De.haf	89893774	M	
De.haf	89893845	M	
De.haf	89894724	M	44H
De.haf	89894877	M	44H
De.haf	89895084	M	44H
De.haf	89895358	M	44H
De.haf	89895368	M	44H
De.haf	89895627	M	
De.haf	89895665	M	44H
De.haf	89895743	bw2a--M	
De.haf	89895935	M	44H
De.haf	89895972	M	
De.haf	89896008	M	44H
De.haf	89896101	M	
De.haf	89896126	M--Y2	44H
De.haf	89896488	W3M	44H
De.haf	89896630	M	
De.haf	89896736	M	44H
De.haf	89896784	M	44H
De.haf	89896865	M	44H
De.haf	89896960	M	
De.haf	89897005	M	
De.haf	89897275	M	44H
De.haf	89897494	M	
De.haf	89897516	M	
De.haf	89897536	M	44H
De.haf	89897735	M	
De.psy	51244013	M	42H
De.psy	51244144	M	42H
De.psy	51244320	M	42H
De.psy	51244347	M	42H
De.psy	51244421	M	42H
De.psy	51244814	MM	42H
De.psy	51244815	MM	42H
De.psy	51244830	M-M-----mM	40H
De.psy	51244831	mM-----m-m	42H
De.psy	51244837	M-M-----mM	42H
De.psy	51244839	M-M-----mM	42H
De.psy	51245128	M	42H
De.psy	51245376	M	42H

Table A.13 (continued)

ID	GI	Gene Neighborhood	Class
De.psy	51245748	MMMM	42H
De.psy	51245749	MMMM	42H
De.psy	51245750	MMMM	42H
De.psy	51245751	MMMM	42H
De.psy	51246110	M	42H
De.psy	51246212	W1M	
De.psy	51246557	M	42H
De.psy	51246845	M	42H
De.des	78355265	M	40H
De.des	78355394	M	40H
De.des	78355416	M	40H
De.des	78355481	M	40H
De.des	78355505	M	40H
De.des	78355750	M	40H
De.des	78356124	MW2	34H
De.des	78356326	M	40H
De.des	78356369	M	40H
De.des	78356670	M	34H
De.des	78356693	M	40H
De.des	78356708	M	40H
De.des	78356795	M--m	40H
De.des	78356798	M--m	
De.des	78357151	MY5A2R3B2	
De.des	78357202	M	34H
De.des	78357454	M	40H
De.des	78357627	M	40H
De.des	78357776	M	40H
De.des	78357855	M	40H
De.des	78357898	M	40H
De.des	78358008	M	40H
De.des	78358060	M	40H
De.des	78358114	M-----v3	40H
De.des	78358286	M	40H
De.des	78358301	M	40H
De.des	78358333	M-M	40H
De.des	78358335	M-M	40H
De.des	78358547	M	40H
De.des	78358600	M	40H
De.vul	46578435	M	40H
De.vul	46578511	M	40H
De.vul	46578587	M	40H
De.vul	46578600	M	40H
De.vul	46578760	M	40H
De.vul	46579005	MW1	34H
De.vul	46579022	M	40H
De.vul	46579059	M	40H

Table A.13 (continued)

ID	GI	Gene Neighborhood	Class
De.vul	46579082	M	40H
De.vul	46579113	M	40H
De.vul	46579163	M	40H
De.vul	46579348	M	40H
De.vul	46579580	M	40H
De.vul	46579811	M	40H
De.vul	46580266	M	40H
De.vul	46580278	M	40H
De.vul	46580293	M	40H
De.vul	46580371	MW3A2	34H
De.vul	46580384	M	40H
De.vul	46580700	M	40H
De.vul	46580714	M	40H
De.vul	46580722	M	40H
De.vul	46580989	M	40H
De.vul	46581142	M	40H
De.vul	46581438	M	40H
De.vul	46581485	M	40H
De.vul	46581557	M	40H
De.vul	46581584	M	40H
Er.car	50119041	M	36H
Er.car	50119052	M	36H
Er.car	50119142	MM	36H
Er.car	50119143	MM	36H
Er.car	50119349	M	36H
Er.car	50119385	M-M	36H
Er.car	50119387	M-M	36H
Er.car	50120044	M	36H
Er.car	50120220	M	36H
Er.car	50120271	M	36H
Er.car	50120444	M	36H
Er.car	50120617	zybrmwa-----M	36H
Er.car	50120625	m-----AWMRBYZ	36H
Er.car	50120707	M	36H
Er.car	50120928	MM	36H
Er.car	50120929	MM	36H
Er.car	50120989	M	36H
Er.car	50121153	M	36H
Er.car	50121245	M	36H
Er.car	50121454	M	36H
Er.car	50121503	MM	36H
Er.car	50121504	MM	36H
Er.car	50121636	M	36H
Er.car	50122167	M	36H
Er.car	50122500	M	36H
Er.car	50122513	M	36H

Table A.13 (continued)

ID	GI	Gene Neighborhood	Class
Er.car	50122536	M	36H
Er.car	50122545	M	36H
Er.car	50122563	M	36H
Er.car	50122759	M	36H
Er.car	50122823	M	36H
Er.car	50123040	M	36H
Er.car	50123252	MMMM	36H
Er.car	50123253	MMMM	36H
Er.car	50123254	MMMM	36H
Er.car	50123255	MMMM	36H
Er.lit	85374015	M	
Er.lit	85375074	AWYBR-M	38H
Es.col	15801723	M	36H
Es.col	15802297	AWMMRBYZ	36H
Es.col	15802298	AWMMRBYZ	36H
Es.col	15803613	M	36H
Es.col	15804929	M	36H
Ge.kau	56418904	M	44H
Ge.kau	56419506	M	44H
Ge.kau	56419555	M	44H
Ge.kau	56419846	M	44H
Ge.kau	56420473	M	
Ge.kau	56421907	M	44H
Ge.met	78221750	M	40H
Ge.met	78221932	M	40H
Ge.met	78222041	M	40H
Ge.met	78222019	M	40H
Ge.met	78222296	A1W1MR2D1B2	36H
Ge.met	78222850	M	
Ge.met	78223625	Y5--Y4A3--MMMW4R4D2B4	34H
Ge.met	78223626	Y5--Y4A3--MMMW4R4D2B4	34H
Ge.met	78223627	Y5--Y4A3--MMMW4R4D2B4	34H
Ge.met	78223681	M	64H
Ge.met	78223905	W5R6W6MA4B6	40H
Ge.met	78224021	M-B7M	40H
Ge.met	78224024	M-B7M	40H
Ge.met	78224134	Mm	64H
Ge.met	78224135	Mm	40H
Ge.met	78224279	M	40H
Ge.met	78224515	M	
Ge.sul	39995508	MM-Y1X2X3	40H
Ge.sul	39995509	MM-Y1X2X3	40H
Ge.sul	39995689	Mm	40H
Ge.sul	39995690	Mm	64H
Ge.sul	39995789	MW2	34H
Ge.sul	39995856	m-----M	40H

Table A.13 (continued)

ID	GI	Gene Neighborhood	Class
Ge.sul	39995862	m-----M	40H
Ge.sul	39995872	M	64H
Ge.sul	39996019	M	40H
Ge.sul	39996038	M	40H
Ge.sul	39996132	MM-MMM-----M	40H
Ge.sul	39996133	MM-MMM-----M	40H
Ge.sul	39996135	MM-MMM-----M	40H
Ge.sul	39996136	MM-MMM-----M	40H
Ge.sul	39996137	MM-MMM-----M	40H
Ge.sul	39996143	MM-MMM-----M	40H
Ge.sul	39996242	MMW3R3D2B2	34H
Ge.sul	39996243	MMW3R3D2B2	34H
Ge.sul	39996389	Y2M-Y3A2---M---MW4MW5-MM	
Ge.sul	39996396	Y2M-Y3A2---M---MW4MW5-MM	34H
Ge.sul	39996400	Y2M-Y3A2---M---MW4MW5-MM	34H
Ge.sul	39996402	Y2M-Y3A2---M---MW4MW5-MM	34H
Ge.sul	39996405	Y2M-Y3A2---M---MW4MW5-MM	34H
Ge.sul	39996406	Y2M-Y3A2---M---MW4MW5-MM	34H
Ge.sul	39996476	M	40H
Ge.sul	39996804	M	
Ge.sul	39997468	M	64H
Ge.sul	39997518	M	
Ge.sul	39997674	MW9	40H
Ge.sul	39997746	M	40H
Ge.sul	39998033	M	40H
Ge.sul	39998246	M	
Ge.sul	39998286	R5MW10Y7A4CD3	44H
Gl.oxy	58039232	M	36H
Gl.oxy	58039792	M	36H
Gl.oxy	58039982	M-Y1AWRBY2	36H
Ha.che	83642995	M	40H
Ha.che	83643349	M----Y1A1W1MMW2-R1D1B1	
Ha.che	83643357	M----Y1A1W1MMW2-R1D1B1	36H
Ha.che	83643358	M----Y1A1W1MMW2-R1D1B1	36H
Ha.che	83643396	M	24H
Ha.che	83643409	M	40H
Ha.che	83643438	Y3Y2W4MR2A2B2W3	40H
Ha.che	83644002	M	40H
Ha.che	83644171	M	40H
Ha.che	83644493	M	40H
Ha.che	83644586	M	40H
Ha.che	83644748	M	40H
Ha.che	83644758	M	40H
Ha.che	83644767	M	40H
Ha.che	83645255	M	40H
Ha.che	83645370	M	40H

Table A.13 (continued)

ID	GI	Gene Neighborhood	Class
Ha.che	83645852	M	24H
Ha.che	83645863	M	40H
Ha.che	83646110	M	40H
Ha.che	83646199	M	40H
Ha.che	83646434	Y5-A3-MW5R3D2B3	34H
Ha.che	83646471	M	40H
Ha.che	83646566	W7R4W6MA4B4	
Ha.che	83647021	M	40H
Ha.che	83647345	m--M	40H
Ha.che	83647348	m--M	36H
Ha.che	83647493	M	
Ha.che	83647533	M	40H
Ha.che	83647556	M	40H
Ha.che	83647634	M	40H
Ha.che	83647717	M	40H
Ha.che	83648471	Y7Y8W10-MA6-W11	
Ha.che	83648708	M	40H
Ha.che	83648937	M	40H
Ha.che	83649336	M	40H
Ha.mar	55376654	M	44H
Ha.mar	55376954	M	44H
Ha.mar	55377103	M	
Ha.mar	55377296	M	
Ha.mar	55377425	M	44H
Ha.mar	55377553	M	44H
Ha.mar	55377770	M	44H
Ha.mar	55377801	M	44H
Ha.mar	55378507	M	
Ha.mar	55378800	M	
Ha.mar	55379092	M	44H
Ha.mar	55379155	M	44H
Ha.mar	55379260	c4---M	
Ha.mar	55379274	M	
Ha.mar	55379349	M	44H
Ha.mar	55379379	M	44H
Ha.mar	55379831	M	44H
Halob	15789618	M	44H
Halob	15789818	M	44H
Halob	15789953	M	44H
Halob	15789962	m---M	44H
Halob	15789966	m---M	44H
Halob	15790077	M	
Halob	15790124	M	
Halob	15790413	M	
Halob	15790447	M	
Halob	15790497	M	

Table A.13 (continued)

ID	GI	Gene Neighborhood	Class
Halob	15790508	M	44H
Halob	15790609	M	44H
Halob	15790664	M	44H
Halob	15790681	Mm--m	44H
Halob	15790682	M--Mm	44H
Halob	15790685	M--Mm	44H
Halob	15790756	M	44H
He.aci	109947910	M	
He.aci	109947957	M	40H
He.aci	109947973	M	28H
He.aci	109948014	M	28H
He.hep	32265980	M	28H
He.hep	32266012	M	40H
He.hep	32266224	M	28H
He.hep	32266390	M	
He.hep	32266470	M	28H
He.hep	32266587	M	28H
He.hep	32266654	M	40H
He.hep	32266809	M	40H
He.hep	32266836	M	28H
He.pyl	15611146	M	28H
He.pyl	15611161	M---m	28H
He.pyl	15611165	M---m	40H
He.pyl	15611613	M	
Id.loi	56459144	M	40H
Id.loi	56459286	M	40H
Id.loi	56459313	M	40H
Id.loi	56459719	M	40H
Id.loi	56460176	M	40H
Id.loi	56460404	M	40H
Id.loi	56460529	M	40H
Id.loi	56460774	M	24H
Id.loi	56460876	M	40H
Id.loi	56460980	M	40H
Id.loi	56461164	M	24H
Id.loi	56461272	M	
Id.loi	56461309	M	
Id.loi	56461335	M	40H
Id.loi	56461443	M	40H
Janna	89053870	M	36H
Janna	89054718	M	
Janna	89054860	M	
Janna	89055851	M	36H
La.int	94972496	M----M	
La.int	94972501	M----M	40H
La.int	94987285	M	40H

Table A.13 (continued)

ID	GI	Gene Neighborhood	Class
Le.int	24212749	M	44H
Le.int	24213376	M-M	44H
Le.int	24213378	M-M	44H
Le.int	24213891	M	44H
Le.int	24213914	M	44H
Le.int	24214946	M	
Le.int	24215126	Y2-A2MW3D1B3	34H
Le.int	24215274	M	44H
Le.int	24215513	M	44H
Le.int	24216205	M	44H
Le.int	24216321	M	
Le.int	24216942	M-----M	
Le.int	24216948	M-----M	
Li.inn	16799806	M	44H
Li.inn	16800875	M	24H
Li.mon	16802765	M	44H
Li.mon	16803739	M	24H
Ma.mag	83309283	M	38H
Ma.mag	83309319	M	38H
Ma.mag	83309344	M	38H
Ma.mag	83309362	M	38H
Ma.mag	83309651	M-M	38H
Ma.mag	83309653	M-M	38H
Ma.mag	83309685	M	38H
Ma.mag	83309827	M	38H
Ma.mag	83309953	M	38H
Ma.mag	83310093	M	38H
Ma.mag	83310316	M	38H
Ma.mag	83310637	M	38H
Ma.mag	83310655	M	
Ma.mag	83310677	M	38H
Ma.mag	83310905	M	38H
Ma.mag	83310967	M	38H
Ma.mag	83311058	M--W2R2W3MB2	38H
Ma.mag	83311064	M--W2R2W3MB2	
Ma.mag	83311083	M	38H
Ma.mag	83311203	M	38H
Ma.mag	83311233	M	38H
Ma.mag	83311295	M	38H
Ma.mag	83311385	M	
Ma.mag	83311432	M	38H
Ma.mag	83311616	M	
Ma.mag	83311751	M---Y5	38H
Ma.mag	83311759	M	38H
Ma.mag	83311869	M	38H
Ma.mag	83311887	M	38H

Table A.13 (continued)

ID	GI	Gene Neighborhood	Class
Ma.mag	83311894	M	38H
Ma.mag	83311920	M----M	38H
Ma.mag	83311925	M----M	
Ma.mag	83312093	M	38H
Ma.mag	83312100	MB4R4A2--W4	40H
Ma.mag	83312153	M	38H
Ma.mag	83312166	M	38H
Ma.mag	83312201	M	38H
Ma.mag	83312291	M	
Ma.mag	83312366	M	38H
Ma.mag	83312413	M	38H
Ma.mag	83312453	M	38H
Ma.mag	83312482	M	38H
Ma.mag	83312566	M	38H
Ma.mag	83312573	M	38H
Ma.mag	83312638	M	38H
Ma.mag	83312653	M	38H
Ma.mag	83312730	M	
Ma.mag	83312789	M	38H
Ma.mag	83312800	Y8---M	38H
Ma.mag	83312819	M	
Ma.mag	83312892	M	38H
Ma.mag	83312903	M	38H
Ma.mag	83312996	M-m-M	38H
Ma.mag	83312998	m-M-m	
Ma.mag	83313000	M-m-M	38H
Ma.mag	83313069	M	38H
Ma.mag	83313090	M	38H
Ma.mag	83313107	M	38H
Ma.mag	83313175	M	38H
Ma.mag	83313255	Y10C--M	38H
Ma.mag	83313263	M	38H
Ma.mag	83313402	M	38H
Me.lot	13488383	MW2RW1AB	40H
Me.bur	91772409	MW-YB1ACDR1	
Me.mar	45357976	M	44H
Me.mar	45358050	M	44H
Me.mar	45358351	M	44H
Me.mar	45358492	WBADMrc1c2y	44H
Me.ace	20088918	d1c1r1a1b1y1--MW1	44H
Me.ace	20091889	MW2-Y2B3A2C2D2R3	44H
Me.ace	73668524	dr1ab1y-MW	44H
Me.maz	21226435	MW1-Y1B1A1C1D1R1	44H
Me.maz	21227431	d2c2r2a2b2y2-MW2	44H
Me.maz	21227760	M	44H
Me.hun	88601330	MM----MW2---W1	44H

Table A.13 (continued)

ID	GI	Gene Neighborhood	Class
Me.hun	88601335	MM----MW2---W1	
Me.hun	88601336	MM----MW2---W1	44H
Me.hun	88601642	M	
Me.hun	88601658	M	
Me.hun	88601665	M	44H
Me.hun	88601798	m--W3MA2	
Me.hun	88601802	a2mw3--M	44H
Me.hun	88601976	M	44H
Me.hun	88602242	Mb3	44H
Me.hun	88602281	W7R3W6MA3B4	40H
Me.hun	88602464	M	
Me.hun	88602637	M	
Me.hun	88602684	m---M	44H
Me.hun	88602688	m---M	
Me.hun	88602698	MW8	
Me.hun	88602913	MW9-----m	
Me.hun	88602920	M-----w9m	
Me.hun	88603103	M	
Me.hun	88603208	M	
Me.hun	88603281	M	
Me.hun	88603317	M	44H
Me.hun	88603791	w14M	44H
Me.hun	88603906	M	
Me.hun	88604177	M---M	44H
Me.hun	88604181	M---M	
Me.hun	88604203	M	44H
Me.fla	91774708	M	36H
Me.fla	91775598	Y1Y2W1MA1	40H
Me.fla	91775923	M	36H
Me.fla	91776285	Y4A2W3MMR1DB1Y3Z	36H
Me.fla	91776286	Y4A2W3MMR1DB1Y3Z	36H
Me.fla	91776506	M	36H
Me.cap	53804220	M	
Me.cap	53804249	M	40H
Mo.the	83589252	M	44H
Mo.the	83589366	M	44H
Mo.the	83589592	MWAB	44H
Mo.the	83590330	M	44H
Mo.the	83590859	M	44H
My.xan	108757399	W10R7MY4W8A6	40H
My.xan	108757523	Y8W2R5W6A2B5M	
My.xan	108757878	A7W1M-R6B2--Y1---b3a3mw3r2w5	
My.xan	108758065	W14W11MA4B7R3	
My.xan	108758139	M	
My.xan	108758538	M	
My.xan	108758729	W5R2W3MA3B3---y1--b2r6-mw1a7	40H

Table A.13 (continued)

ID	GI	Gene Neighborhood	Class
My.xan	108759887	M	
My.xan	108760175	M	
My.xan	108761207	W7-MMA5-B6R4	
My.xan	108761607	M	
My.xan	108761813	M	
My.xan	108762109	W7-MMA5-B6R4	40H
My.xan	108762134	M	40H
My.xan	108762203	M	64H
My.xan	108762581	M-W13Y3A8C	44H
My.xan	108763249	M	
My.xan	108763445	M	
My.xan	108763634	M	
My.xan	108763683	M	
My.xan	108763763	M	
Na.pha	76800685	M	
Na.pha	76801312	M	
Na.pha	76801394	M	
Na.pha	76801471	mM	
Na.pha	76801472	mM	44H
Na.pha	76801509	M	
Na.pha	76801527	M	44H
Na.pha	76801578	M	
Na.pha	76801732	rabM	44H
Na.pha	76801947	M	
Na.pha	76802022	M	
Na.pha	76802088	M	
Na.pha	76802201	M-----M-c2	44H
Na.pha	76802207	M-----M-c2	44H
Na.pha	76802246	M	
Na.pha	76802325	M	
Na.pha	76802456	M	44H
Na.pha	76803046	M	44H
Na.pha	76803424	M	44H
Ni.ham	92115981	M	
Ni.ham	92116384	M	38H
Ni.ham	92116624	M	38H
Ni.ham	92116659	M	38H
Ni.win	75674728	AWY1BR-----M	38H
Ni.win	75677177	M	38H
Ni.oce	77163666	Y2Y1W2MR1AB1W1	40H
Ni.eur	30249233	MA1	40H
Ni.eur	30249815	A2W3MM-RDB	36H
Ni.eur	30249816	A2W3MM-RDB	36H
Ni.mul	82701466	W1RW2MAB	40H
Nosto	17228422	Y2Y1W1MA1	40H
Nosto	17228564	Y4Y3W2MA2	40H

Table A.13 (continued)

ID	GI	Gene Neighborhood	Class
Nosto	17229654	Y6Y5W3MA3	40H
No.aro	87198874	M	
No.aro	87199451	M	
No.aro	87200834	M	
No.aro	87201060	M	
Oc.ihe	23097989	M	44H
Oc.ihe	23098580	M	44H
Oc.ihe	23098712	M	24H
Oc.ihe	23099130	M	44H
Oc.ihe	23099557	M	44H
Oc.ihe	23100163	M	
Oc.ihe	23100184	M	44H
Oc.ihe	23100442	M	44H
Pe.car	77917647	MW1A1	44H
Pe.car	77917670	M	40H
Pe.car	77918054	M	36H
Pe.car	77918160	Mw2	36H
Pe.car	77918199	M	36H
Pe.car	77918580	M	36H
Pe.car	77918940	M	36H
Pe.car	77918958	M	36H
Pe.car	77918974	M	36H
Pe.car	77919581	M	
Pe.car	77919897	M	36H
Pe.car	77919958	MW5	36H
Pe.car	77920061	M	
Pe.car	77920200	M	36H
Pe.car	77920557	M	36H
Ph.pro	54301708	M	40H
Ph.pro	54301780	M	40H
Ph.pro	54301884	M	40H
Ph.pro	54301900	M	40H
Ph.pro	54302045	M-m	40H
Ph.pro	54302047	M-m	40H
Ph.pro	54302061	M	40H
Ph.pro	54302087	M	40H
Ph.pro	54302188	mM	40H
Ph.pro	54302189	mM	40H
Ph.pro	54302333	M	40H
Ph.pro	54303005	M	40H
Ph.pro	54303417	M	40H
Ph.pro	54303523	M-Mm	40H
Ph.pro	54303525	M-Mm	40H
Ph.pro	54303526	Mm-m	40H
Ph.pro	54303603	M	40H
Ph.pro	54307660	M	40H

Table A.13 (continued)

ID	GI	Gene Neighborhood	Class
Ph.pro	54307966	MY2-A1-MR1B1	
Ph.pro	54307971	MY2-A1-MR1B1	34H
Ph.pro	54308182	M	40H
Ph.pro	54308308	M	
Ph.pro	54308329	M	40H
Ph.pro	54308447	M	40H
Ph.pro	54308562	M	40H
Ph.pro	54308588	M	40H
Ph.pro	54308659	M	40H
Ph.pro	54309188	M	40H
Ph.pro	54309203	M	40H
Ph.pro	54309261	M	40H
Ph.pro	54309325	M	40H
Ph.pro	54309371	M	40H
Ph.pro	54309471	M	40H
Ph.pro	54309570	M	40H
Ph.pro	54309987	MM	40H
Ph.pro	54309988	MM	40H
Ph.pro	54310281	M	40H
Ph.pro	54310442	M	40H
Ph.pro	54310451	M	40H
Ph.lum	37525782	AWMMRBYZ	36H
Ph.lum	37525783	AWMMRBYZ	36H
Polar	91787037	Y1Y2W1MA1	40H
Polar	91788328	W2RW3--MMMA2B	40H
Polar	91788329	W2RW3--MMMA2B	40H
Polar	91788330	W2RW3--MMMA2B	40H
Ps.atl	109896399	M	40H
Ps.atl	109896474	M	40H
Ps.atl	109896848	M	40H
Ps.atl	109896932	M	40H
Ps.atl	109898564	M	40H
Ps.atl	109899729	M	40H
Ps.atl	109899765	M	40H
Ps.atl	109899778	M	24H
Ps.atl	109900479	M-M	40H
Ps.atl	109900481	M-M	40H
Ps.hal	77359361	M	40H
Ps.hal	77359662	M	40H
Ps.hal	77359797	M	40H
Ps.hal	77359883	M	40H
Ps.hal	77361303	M	40H
Ps.hal	77361399	M	
Ps.hal	77361553	M	40H
Ps.hal	77361566	M	
Ps.hal	77361648	M	40H

Table A.13 (continued)

ID	GI	Gene Neighborhood	Class
Ps.hal	77361657	M	40H
Ps.hal	77362042	M	40H
Ps.hal	77362101	M	40H
Ps.hal	77362190	M	
Ps.hal	77362427	M	40H
Ps.aer	15595374	MY1A1W1MR1DB1	36H
Ps.aer	15595378	MY1A1W1MR1DB1	
Ps.aer	15595608	Y2Y3W2MR2A2B2W3	40H
Ps.aer	15596448	M	40H
Ps.aer	15596620	M	24H
Ps.aer	15596758	M	40H
Ps.aer	15596805	M	40H
Ps.aer	15596843	M	40H
Ps.aer	15597126	M	24H
Ps.aer	15597757	M	40H
Ps.aer	15597769	M	40H
Ps.aer	15597848	m-M	40H
Ps.aer	15597850	m-M	40H
Ps.aer	15597984	M	40H
Ps.aer	15598063	M	40H
Ps.aer	15598116	M	40H
Ps.aer	15598903	MW7R4W6A4B4	40H
Ps.aer	15599486	M	
Ps.aer	15599503	M-MM	40H
Ps.aer	15599505	M-MM	40H
Ps.aer	15599506	M-MM	40H
Ps.aer	15599716	M	40H
Ps.aer	15599829	M	40H
Ps.aer	15600037	M	40H
Ps.aer	15600108	M	40H
Ps.aer	15600265	M	40H
Ps.ent	104779920	M	40H
Ps.ent	104779960	M	40H
Ps.ent	104780126	M	40H
Ps.ent	104780144	M	24H
Ps.ent	104780151	M	40H
Ps.ent	104780168	M---v1	40H
Ps.ent	104780365	M	40H
Ps.ent	104780451	MW1R2W2A1B1	40H
Ps.ent	104780695	M	40H
Ps.ent	104780988	M	40H
Ps.ent	104780998	M	40H
Ps.ent	104781023	M	40H
Ps.ent	104781177	M	40H
Ps.ent	104781535	M	24H
Ps.ent	104781861	M	40H

Table A.13 (continued)

ID	GI	Gene Neighborhood	Class
Ps.ent	104782002	M	40H
Ps.ent	104782061	M	40H
Ps.ent	104782570	M	
Ps.ent	104782647	M	40H
Ps.ent	104782755	M	40H
Ps.ent	104782763	M	40H
Ps.ent	104782926	M	40H
Ps.ent	104783055	M	40H
Ps.ent	104783791	M	40H
Ps.ent	104783866	M	40H
Ps.ent	104783970	Y3Y2W6MA3W5	40H
Ps.ent	104784000	MM	40H
Ps.ent	104784001	MM	40H
Ps.ent	104784019	M	40H
Ps.ent	104784077	m--M	40H
Ps.ent	104784080	m--M	40H
Ps.ent	104784305	M	40H
Ps.flu	70728511	MW1R1W2A1B1	40H
Ps.flu	70729125	M	40H
Ps.flu	70729147	M	40H
Ps.flu	70729283	M	40H
Ps.flu	70729309	M	40H
Ps.flu	70729551	M	40H
Ps.flu	70729680	M	40H
Ps.flu	70729696	M	40H
Ps.flu	70729775	M	40H
Ps.flu	70729905	M	40H
Ps.flu	70729965	M	
Ps.flu	70730262	M	40H
Ps.flu	70730480	M	40H
Ps.flu	70730683	M	40H
Ps.flu	70730719	M	40H
Ps.flu	70731367	m-M	40H
Ps.flu	70731369	m-M	40H
Ps.flu	70731638	M	40H
Ps.flu	70731875	M	24H
Ps.flu	70731886	M	40H
Ps.flu	70731932	M	24H
Ps.flu	70731972	M	40H
Ps.flu	70732109	M	40H
Ps.flu	70732369	M---v3	40H
Ps.flu	70732442	M	40H
Ps.flu	70732526	MM	24H
Ps.flu	70732527	MM	24H
Ps.flu	70732694	M	40H
Ps.flu	70733109	Y3Y2W6MA3W5	40H

Table A.13 (continued)

ID	GI	Gene Neighborhood	Class
Ps.flu	70733232	M	40H
Ps.flu	70733311	M	40H
Ps.flu	70733635	M	40H
Ps.flu	70733899	M	40H
Ps.flu	70733922	M	40H
Ps.flu	70733993	m--MM	40H
Ps.flu	70733994	m--MM	40H
Ps.flu	70733997	mm--M	40H
Ps.flu	70734076	M	40H
Ps.flu	70734176	M	40H
Ps.flu	70734209	M	40H
Ps.flu	70734275	M	40H
Ps.flu	70734347	M	40H
Ps.put	26987059	m--M	40H
Ps.put	26987062	m--M	40H
Ps.put	26987300	M	40H
Ps.put	26987322	M	40H
Ps.put	26987515	M	24H
Ps.put	26987963	M	40H
Ps.put	26988105	M	40H
Ps.put	26988221	MW1R1W2A1B1	40H
Ps.put	26988549	M	40H
Ps.put	26988667	M	40H
Ps.put	26988835	M	40H
Ps.put	26988844	M	40H
Ps.put	26988973	M	40H
Ps.put	26988981	M	40H
Ps.put	26989034	M	40H
Ps.put	26989362	M	40H
Ps.put	26989542	M	40H
Ps.put	26989580	M	40H
Ps.put	26990127	M	24H
Ps.put	26990269	M	40H
Ps.put	26990655	M	
Ps.put	26991206	M	40H
Ps.put	26991342	M	40H
Ps.put	26991566	M	40H
Ps.put	26991666	Y3Y2W6MA3W5	40H
Ps.put	26991696	MM	40H
Ps.put	26991697	MM	40H
Ps.syr	28867357	M	40H
Ps.syr	28867495	M	40H
Ps.syr	28867696	M	40H
Ps.syr	28868132	MY1-A1MW1R1DB1	34H
Ps.syr	28868136	MY1-A1MW1R1DB1	
Ps.syr	28868215	M	40H

Table A.13 (continued)

ID	GI	Gene Neighborhood	Class
Ps.syr	28868227	M	24H
Ps.syr	28868276	m--m-M	40H
Ps.syr	28868278	m-M--M	40H
Ps.syr	28868281	m-M--M	40H
Ps.syr	28868542	M	40H
Ps.syr	28868700	MW2R2W3A2B2	40H
Ps.syr	28868854	M	40H
Ps.syr	28869218	M	40H
Ps.syr	28869455	M	40H
Ps.syr	28869636	MW6-----m	
Ps.syr	28869643	M-----w6m	40H
Ps.syr	28869665	M--M----M	40H
Ps.syr	28869668	M--M----M	40H
Ps.syr	28869673	M--M----M	40H
Ps.syr	28869704	M	40H
Ps.syr	28869719	M	40H
Ps.syr	28869806	M	40H
Ps.syr	28870064	M	24H
Ps.syr	28870175	M	40H
Ps.syr	28870271	M	40H
Ps.syr	28870402	M	40H
Ps.syr	28870443	M	40H
Ps.syr	28870455	M	40H
Ps.syr	28870543	M	40H
Ps.syr	28870641	M	
Ps.syr	28870737	M--m	40H
Ps.syr	28870740	M--m	40H
Ps.syr	28870835	m----M	40H
Ps.syr	28870840	m----M	40H
Ps.syr	28870854	M	40H
Ps.syr	28870908	M	24H
Ps.syr	28871665	M	
Ps.syr	28871675	M	40H
Ps.syr	28871756	M	40H
Ps.syr	28871902	M	24H
Ps.syr	28872050	M	40H
Ps.syr	28872145	Y4Y3W8MA4W7	40H
Ps.syr	28872271	MM	40H
Ps.syr	28872272	MM	40H
Ps.syr	28872463	M	24H
Ps.syr	28872658	MM	40H
Ps.syr	28872659	MM	40H
Ps.syr	28872674	M	40H
Ps.arc	71066370	Y2Y1WMRA	40H
Ps.cry	93006922	Y3Y2WMRAB	40H
Py.aby	14520639	M	44H

Table A.13 (continued)

ID	GI	Gene Neighborhood	Class
Py.aby	14521748	wmRYBAC2C1DM	44H
Py.aby	14521756	mdc1c2abyrMW	
Py.aby	14521788	M	44H
Py.aby	14521956	M	
Py.hor	14590358	M	44H
Py.hor	14590391	mdc2c1abyr-MW	
Py.hor	14590400	wm-RYBAC1C2DM	44H
Py.hor	14591598	M	44H
Py.hor	14591707	M	
Ra.eut	72384050	M	36H
Ra.eut	72384125	M	36H
Ra.eut	73537488	MW2R1W1A1B1	40H
Ra.eut	73537560	M-----M	36H
Ra.eut	73537566	M-----M	36H
Ra.eut	73537643	M	36H
Ra.eut	73537834	M	36H
Ra.eut	73538912	M	36H
Ra.eut	73538934	M	36H
Ra.eut	73538961	M	36H
Ra.eut	73539053	M	36H
Ra.eut	73539369	M	36H
Ra.eut	73539419	M-Y1W3MM-----A2W4R4DB4Y2Z	36H
Ra.eut	73539423	M-Y1W3MM-----A2W4R4DB4Y2Z	36H
Ra.eut	73539424	M-Y1W3MM-----A2W4R4DB4Y2Z	36H
Ra.eut	73539688	M	36H
Ra.eut	73539742	M	36H
Ra.eut	73540475	M	
Ra.eut	73541374	M	36H
Ra.eut	73541645	M-Y3	
Ra.eut	73541708	M	36H
Ra.eut	73542138	M	36H
Ra.eut	73542305	Y5Y4W5MA3	40H
Ra.met	94309105	M	36H
Ra.met	94309617	Y1Y2W1MA1	40H
Ra.met	94310633	M	
Ra.met	94312610	M-Y3W2MM-----A2W3R2DB1Y4Z	36H
Ra.met	94312614	M-Y3W2MM-----A2W3R2DB1Y4Z	36H
Ra.met	94312615	M-Y3W2MM-----A2W3R2DB1Y4Z	36H
Ra.met	94312655	M	36H
Ra.met	94312901	MW5R3W4A3B2	40H
Ra.met	94313112	M	36H
Ra.met	94313122	M	36H
Ra.met	94313649	M	36H
Ra.met	94314169	M	36H
Ra.met	94314225	M	36H
Ra.met	94314531	M	36H

Table A.13 (continued)

ID	GI	Gene Neighborhood	Class
Ra.met	94314560	M	36H
Ra.met	94314854	M	36H
Ra.sol	17545325	M	36H
Ra.sol	17545390	Y1Y2W1MA1	40H
Ra.sol	17545874	MM	36H
Ra.sol	17545875	MM	36H
Ra.sol	17545953	M	36H
Ra.sol	17546179	M	36H
Ra.sol	17546613	M	36H
Ra.sol	17546669	M	36H
Ra.sol	17547518	M	36H
Ra.sol	17547855	M	36H
Ra.sol	17548024	M	36H
Ra.sol	17548129	M	36H
Ra.sol	17548476	M	36H
Ra.sol	17548524	M	36H
Ra.sol	17548728	M	36H
Ra.sol	17549061	M	36H
Ra.sol	17549202	M	
Ra.sol	17549248	M	36H
Ra.sol	17549320	M	36H
Ra.sol	17549430	M	36H
Ra.sol	17549445	M	36H
Ra.sol	17549582	M	36H
Ra.sol	17549625	Y5A2W2MR1DBY4Z2	36H
Rh.etl	86356065	m--M	36H
Rh.etl	86356068	m--M	36H
Rh.etl	86356097	M	36H
Rh.etl	86356185	M	36H
Rh.etl	86356287	M-Y1A1W1R1B1Y2D	36H
Rh.etl	86356359	MM	36H
Rh.etl	86356360	MM	36H
Rh.etl	86356534	M	36H
Rh.etl	86356558	M	36H
Rh.etl	86356620	M	36H
Rh.etl	86356801	M	36H
Rh.etl	86356894	M	
Rh.etl	86358156	M	36H
Rh.etl	86358433	MW2	36H
Rh.etl	86359068	M	36H
Rh.etl	86359109	Y4A2W4MMMW3R2B2	34H
Rh.etl	86359110	Y4A2W4MMMW3R2B2	34H
Rh.etl	86359111	Y4A2W4MMMW3R2B2	34H
Rh.etl	86360197	M	36H
Rh.etl	86360909	M	36H
Rh.etl	86360980	M	36H

Table A.13 (continued)

ID	GI	Gene Neighborhood	Class
Rh.etl	86361029	M	36H
Rh.etl	86361092	M	36H
Rh.etl	86361181	M	36H
Rh.sph	77462998	Y4M-MD-Y3A2W3R3Y2	36H
Rh.sph	77463000	Y4M-MD-Y3A2W3R3Y2	34H
Rh.sph	77463615	A3R4B3W4-MY5A4	
Rh.sph	77463786	M	
Rh.sph	77463802	M	34H
Rh.sph	77465097	M	34H
Rh.sph	77465121	M	34H
Rh.sph	77465309	Y6M	34H
Rh.sph	77465422	M	34H
Rh.sph	77465434	M	34H
Rh.sph	77465716	M	34H
Rh.fer	89898941	M	36H
Rh.fer	89899706	MY2-A2MW3R2D2B2	
Rh.fer	89899710	MY2-A2MW3R2D2B2	34H
Rh.fer	89899732	W4M	36H
Rh.fer	89899794	W5M	36H
Rh.fer	89900168	Y3Y4W6MA3	40H
Rh.fer	89900273	M---w7	36H
Rh.fer	89900814	M	36H
Rh.fer	89901008	M	36H
Rh.fer	89901119	M	36H
Rh.fer	89901127	b4-b3/r3m---M	36H
Rh.fer	89901131	m---MB3/R3-B4	36H
Rh.fer	89901258	M	36H
Rh.fer	89901284	M	36H
Rh.fer	89901522	M	36H
Rh.fer	89901561	M	36H
Rh.fer	89901794	M	36H
Rh.fer	89901888	M	36H
Rh.fer	89901906	M	36H
Rh.fer	89902055	M	36H
Rh.fer	89902217	M	36H
Rh.fer	89902372	MM	36H
Rh.fer	89902373	MM	36H
Rh.fer	89902696	M	36H
Rh.pal	39933216	Y1A1W2W1MR1B1	34H
Rh.pal	39933311	M	38H
Rh.pal	39933508	M	24H
Rh.pal	39933933	M	38H
Rh.pal	39934171	M	
Rh.pal	39934710	M	38H
Rh.pal	39934745	MA3W4R3	38H
Rh.pal	39934891	M	38H

Table A.13 (continued)

ID	GI	Gene Neighborhood	Class
Rh.pal	39934919	M	38H
Rh.pal	39934953	M	24H
Rh.pal	39935000	M	38H
Rh.pal	39935545	M	38H
Rh.pal	39936248	M	38H
Rh.pal	39936608	M	38H
Rh.pal	39936660	M	38H
Rh.pal	39936811	M	38H
Rh.pal	39937262	M	38H
Rh.pal	39937362	M---MM---M	38H
Rh.pal	39937366	M---MM---M	38H
Rh.pal	39937367	M---MM---M	38H
Rh.pal	39937371	M---MM---M	38H
Rh.pal	39937509	M	38H
Rh.pal	39937541	M-M	38H
Rh.pal	39937543	M-M	38H
Rh.pal	39937696	MM	38H
Rh.pal	39937697	MM	38H
Rh.pal	39937721	M	38H
Rh.pal	39937742	M	38H
Rh.pal	39937749	M	38H
Rh.rub	83591460	M	38H
Rh.rub	83591470	M	
Rh.rub	83591498	M	38H
Rh.rub	83591602	M	38H
Rh.rub	83591781	M	38H
Rh.rub	83591907	M	38H
Rh.rub	83591915	M	38H
Rh.rub	83591971	M	38H
Rh.rub	83592057	M	38H
Rh.rub	83592094	M	38H
Rh.rub	83592389	M	38H
Rh.rub	83592445	M	38H
Rh.rub	83592496	M	38H
Rh.rub	83592513	M	38H
Rh.rub	83592533	M	38H
Rh.rub	83592629	M	38H
Rh.rub	83592737	Y2A2W1MW2MMR2B2Dm	34H
Rh.rub	83592739	Y2A2W1MW2MMR2B2Dm	34H
Rh.rub	83592740	Y2A2W1MW2MMR2B2Dm	34H
Rh.rub	83592744	Mdb2r2mmw2mw1a2y2	38H
Rh.rub	83592889	M	
Rh.rub	83592918	MM	38H
Rh.rub	83592919	MM	38H
Rh.rub	83593047	M	38H
Rh.rub	83593137	M	38H

Table A.13 (continued)

ID	GI	Gene Neighborhood	Class
Rh.rub	83593236	M	38H
Rh.rub	83593358	M	38H
Rh.rub	83593411	M-M	38H
Rh.rub	83593413	M-M	38H
Rh.rub	83593477	M---M	38H
Rh.rub	83593481	M---M	38H
Rh.rub	83593656	MW3B4-R4Y3A3--Y4--M	
Rh.rub	83593668	MW3B4-R4Y3A3--Y4--M	
Rh.rub	83593701	M	38H
Rh.rub	83593795	M	38H
Rh.rub	83593819	M	38H
Rh.rub	83593881	M	38H
Rh.rub	83593889	M	38H
Rh.rub	83594089	M	38H
Rh.rub	83594106	M	38H
Rh.rub	83594367	M	38H
Rh.rub	83594571	M	38H
Rh.rub	83594794	M	38H
Rh.rub	83594825	M	38H
Rh.rub	83594842	M	38H
Rh.rub	83594892	M	38H
Rh.rub	83594994	M	38H
Sa.deg	90019730	M	40H
Sa.deg	90019992	M	40H
Sa.deg	90020040	C-M	40H
Sa.deg	90020169	Y2Y1W1MA1	40H
Sa.deg	90020318	M	40H
Sa.deg	90020755	M	40H
Sa.deg	90021036	M	
Sa.deg	90021961	M	40H
Sa.deg	90022747	mm--Y5A3W4MR3DB3	36H
Sa.deg	90022753	b3dr3mw4a3y5--MM	
Sa.deg	90022754	b3dr3mw4a3y5--MM	36H
Sa.deg	90023065	M	40H
Sa.deg	90023271	Y7Y6W6MR4A4B4W5	40H
Sa.rub	83814134	M	
Sa.rub	83814438	M	40H
Sa.rub	83814592	M	40H
Sa.rub	83814728	MM	40H
Sa.rub	83814750	M	40H
Sa.rub	83814835	M	
Sa.rub	83815037	M	
Sa.rub	83815042	MM	40H
Sa.rub	83815551	M	40H
Sa.rub	83815947	WR2Y-AB1-M-M-M	40H
Sa.rub	83816322	WR2Y-AB1-M-M-M	40H

Table A.13 (continued)

ID	GI	Gene Neighborhood	Class
Sa.rub	83816473	WR2Y-AB1-M-M-M	
Sa.rub	83816813	M	40H
Sa.ent	16760230	M	36H
Sa.ent	16760868	AWMRBYZ	36H
Sa.ent	16761916	M	
Sa.ent	16761989	MM	36H
Sa.ent	16761990	MM	36H
Sa.ent	16762727	M	36H
Sh.den	91791468	M	40H
Sh.den	91792074	M	40H
Sh.den	91792221	M	40H
Sh.den	91792271	M	40H
Sh.den	91792451	M	40H
Sh.den	91792956	MM	40H
Sh.den	91792957	MM	40H
Sh.den	91793318	M	40H
Sh.den	91793591	M	40H
Sh.den	91793787	M	40H
Sh.den	91793989	M	40H
Sh.den	91794242	M	40H
Sh.den	91794305	M	40H
Sh.den	91794360	M	40H
Sh.den	91794416	M	40H
Sh.den	91794449	M	40H
Sh.den	91794641	MY2-A2MW3R2DB2	
Sh.den	91794645	MY2-A2MW3R2DB2	34H
Sh.den	91794887	M	
Sh.one	24372094	M	40H
Sh.one	24372177	M	40H
Sh.one	24372574	M	40H
Sh.one	24372641	M	40H
Sh.one	24372727	M	
Sh.one	24372859	M	40H
Sh.one	24372963	M	40H
Sh.one	24373012	M	40H
Sh.one	24373643	M	40H
Sh.one	24373677	M--Y1A1W1MR1D1B1	
Sh.one	24373683	M--Y1A1W1MR1D1B1	36H
Sh.one	24373793	x1M	40H
Sh.one	24373870	Y2--MW2R2D2B2	34H
Sh.one	24374574	M	40H
Sh.one	24374793	M	40H
Sh.one	24374914	M	40H
Sh.one	24375014	M	
Sh.one	24375141	M	40H
Sh.one	24375328	M	40H

Table A.13 (continued)

ID	GI	Gene Neighborhood	Class
Sh.one	24375378	M	40H
Sh.one	24375540	M	40H
Sh.one	24375932	M	40H
Sh.one	24375944	M	40H
Sh.one	24376031	M	40H
Sh.one	24376108	M	
Sh.one	24376324	M	40H
Sh.one	50261353	M	40H
Sh.boy	82545327	M	
Sh.boy	82546708	M	
Sh.dys	82777019	M	
Sh.dys	82778401	M	
Sh.fle	30062892	M	36H
Sh.fle	30063337	AWMMRBYZ	36H
Sh.fle	30063338	AWMMRBYZ	36H
Sh.fle	30065597	M	36H
Sh.son	74311764	AWMMRBYZ	36H
Sh.son	74311765	AWMMRBYZ	36H
Sh.son	74312227	M	36H
Sh.son	74313607	M	36H
Silic	99078155	M-M	36H
Silic	99078157	M-M	36H
Silic	99078172	M	36H
Silic	99078187	y1r1w1ay2-MB1D1	36H
Silic	99078220	MR2D2W2	
Silic	99080200	M	36H
Silic	99080274	M	36H
Silic	99080317	M	36H
Silic	99080680	M	36H
Silic	99080825	M	36H
Silic	99080868	MB3/R4	36H
Silic	99080924	M	36H
Silic	99081224	M	36H
Silic	99081328	M	36H
Silic	99081817	M	36H
Silic	99081842	M	36H
Silic	99082293	M	36H
Silic	99082487	M	36H
Silic	99082777	M	36H
Si.mel	15964169	M	36H
Si.mel	15964200	M	36H
Si.mel	15964389	M-Y1A1W1R1B1Y2D	36H
Si.mel	15964462	M	36H
Si.mel	15964630	M	36H
Si.mel	15965569	M	
Si.mel	15965650	M	36H

Table A.13 (continued)

ID	GI	Gene Neighborhood	Class
Si.mel	15965905	MW2	36H
Si.mel	15966759	W3--M	36H
Si.mel	16263300	R2W4MA2B2	40H
Sp.ala	103485730	M	
Sp.ala	103486049	M	
Sp.ala	103487126	M	
Sy.the	51891501	M	52H
Sy.the	51891769	M	52H
Sy.the	51891814	M	52H
Sy.the	51891831	M	52H
Sy.the	51891861	M	52H
Sy.the	51891964	M	
Sy.the	51892086	M	52H
Sy.the	51892220	MM	52H
Sy.the	51892221	MM	52H
Sy.the	51892559	M	
Sy.the	51893012	M	
Sy.the	51893903	M	
Sy.the	51894102	M	52H
Sy.elo	56750540	W1MA1	40H
Sy.elo	56750691	Y3Y2W3MA2W2	40H
Synco2	86605934	M	
Synco2	86606796	Y3Y2W1---MA1	40H
Synco2	86607345	Y4Y5Y6W2MA2	40H
Syncy	16329620	Y1Y2W1M	40H
Syncy	16329791	Y4Y3W2MA1	40H
Syncy	16331987	Y6Y5W4MMA3W3	
Syncy	16331988	Y6Y5W4MMA3W3	40H
Sy.aci	85858536	W1R1W2MA1B1	
Sy.aci	85859025	x1y2a2b2d2-y1-d1r2m---W3--M	36H
Sy.aci	85859032	m--w3---MR2D1-Y1-D2B2A2Y2X1	36H
Sy.aci	85860312	M	
Th.ten	20806957	M	44H
Th.ten	20806973	M	
Th.ten	20807139	M	
Th.ten	20807169	M	44H
Th.ten	20807189	W1-M	44H
Th.ten	20807243	M	44H
Th.ten	20807512	MW2B1-R1Y1A1	44H
Th.ten	20808196	M	44H
Th.ten	20808606	M	44H
Th.kod	57640091	M	44H
Th.kod	57640565	dmc2c1aabyrMW	
Th.kod	57640573	wmRYBAAC1C2MD	44H
Th.kod	57641541	M	
Th.kod	57642082	M	44H

Table A.13 (continued)

ID	GI	Gene Neighborhood	Class
Th.elo	22297891	Y1Y2W1MA1	40H
Th.elo	22298112	Y4Y3W3MA2W2	40H
Th.elo	22298566	Y6Y5W4MA3	40H
Th.mar	15642789	M	
Th.mar	15642798	M	44H
Th.mar	15643195	M	44H
Th.mar	15643680	M	44H
Th.mar	15643900	m--M	44H
Th.mar	15643903	m--M	44H
Th.mar	15644179	M	44H
Tb.den	74316678	M	36H
Tb.den	74317636	Y2A1W1M-MMMRDBY1Z	36H
Tb.den	74317637	Y2A1W1M-MMMRDBY1Z	36H
Tb.den	74317638	Y2A1W1M-MMMRDBY1Z	36H
Tb.den	74317640	Y2A1W1M-MMMRDBY1Z	36H
Tb.den	74318424	M	36H
Tb.den	74318567	Y4Y3W2MA2	40H
Tm.cru	78484373	M	36H
Tm.cru	78484423	M	36H
Tm.cru	78484706	M	36H
Tm.cru	78484898	M	36H
Tm.cru	78484915	M---M	40H
Tm.cru	78484919	M---M	36H
Tm.cru	78485104	Y1ZA1W2--y2w3-RBM	
Tm.cru	78485122	M	36H
Tm.cru	78485738	M	36H
Tm.cru	78485777	M	
Tm.cru	78486236	M	36H
Tm.cru	78486323	M	36H
Tm.cru	78486338	M----m	40H
Tm.cru	78486343	M----m	36H
Tm.den	78776244	M	
Tm.den	78776410	M	
Tm.den	78776470	M	28H
Tm.den	78776484	M	
Tm.den	78777170	Y1W1M-A1R1-DB1-Z1M-M----mM	36H
Tm.den	78777179	Y1W1M-A1R1-DB1-Z1M-M----mM	
Tm.den	78777181	Y1W1M-A1R1-DB1-Z1M-M----mM	
Tm.den	78777186	mM----m-mz1-b1d-r1a1-mw1y1	40H
Tm.den	78777187	Y1W1M-A1R1-DB1-Z1M-M----mM	
Tm.den	78777376	M	
Tm.den	78777693	M	
Tm.den	78777850	M	
Tm.den	78778126	M	40H
Tr.den	42525591	M	48H
Tr.den	42525687	M	

Table A.13 (continued)

ID	GI	Gene Neighborhood	Class
Tr.den	42525696	M--M	48H
Tr.den	42525699	M--M	48H
Tr.den	42525861	M-M	48H
Tr.den	42525863	M-M	48H
Tr.den	42526000	M	48H
Tr.den	42526135	M	48H
Tr.den	42526156	M	48H
Tr.den	42526362	M	48H
Tr.den	42526520	M	48H
Tr.den	42526564	M	48H
Tr.den	42526793	M	48H
Tr.den	42526894	M	48H
Tr.den	42527645	M	48H
Tr.den	42527772	M	48H
Tr.den	42527996	M	
Tr.den	42528049	M	48H
Tr.den	42528085	M	
Tr.den	42528283	M	48H
Tr.pal	15639034	M	
Tr.pal	15639479	M	48H
Tr.pal	15639626	MM	48H
Tr.pal	15639627	MM	48H
Vi.cho	15600779	M	40H
Vi.cho	15600802	M	40H
Vi.cho	15600839	M	40H
Vi.cho	15600946	M	40H
Vi.cho	15600989	M	40H
Vi.cho	15601036	M	40H
Vi.cho	15601416	m----M	40H
Vi.cho	15601421	m----M	40H
Vi.cho	15601528	M	40H
Vi.cho	15601619	M	24H
Vi.cho	15601660	M	40H
Vi.cho	15601677	M	40H
Vi.cho	15601727	M----M	40H
Vi.cho	15601732	M----M	40H
Vi.cho	15601741	M	40H
Vi.cho	15601786	M	40H
Vi.cho	15601807	M	40H
Vi.cho	15601820	M	40H
Vi.cho	15601838	Y1A1W2W1MR1DM	
Vi.cho	15601841	Y1A1W2W1MR1DM	36H
Vi.cho	15640130	M	36H
Vi.cho	15640246	M	40H
Vi.cho	15640311	M	40H
Vi.cho	15640476	M	40H

Table A.13 (continued)

ID	GI	Gene Neighborhood	Class
Vi.cho	15640536	M-M	40H
Vi.cho	15640538	M-M	40H
Vi.cho	15640842	M	40H
Vi.cho	15640857	M	40H
Vi.cho	15641261	M	40H
Vi.cho	15641302	M	40H
Vi.cho	15641311	M	40H
Vi.cho	15641325	M	40H
Vi.cho	15641406	mm-MW3B1-R2Y3A2-M	
Vi.cho	15641414	mm-MW3B1-R2Y3A2-M	
Vi.cho	15641416	m-a2y3r2-b1w3m-MM	40H
Vi.cho	15641417	m-a2y3r2-b1w3m-MM	24H
Vi.cho	15641424	M	40H
Vi.cho	15641543	M	40H
Vi.cho	15641648	M	40H
Vi.cho	15641861	M	40H
Vi.cho	15641870	M	40H
Vi.cho	15641900	M	40H
Vi.cho	15641969	M	40H
Vi.cho	15642160	M	40H
Vi.cho	15642435	M	40H
Vi.fis	59711305	M	40H
Vi.fis	59711384	M	40H
Vi.fis	59711434	M	40H
Vi.fis	59711479	M	
Vi.fis	59711594	M	40H
Vi.fis	59711698	MM	40H
Vi.fis	59711699	MM	40H
Vi.fis	59711724	M	40H
Vi.fis	59711740	M----M	40H
Vi.fis	59711745	M----M	40H
Vi.fis	59711976	M	40H
Vi.fis	59712110	MM	40H
Vi.fis	59712111	MM	40H
Vi.fis	59712225	M	40H
Vi.fis	59712259	M	40H
Vi.fis	59712385	M	40H
Vi.fis	59712396	M	40H
Vi.fis	59712649	M	40H
Vi.fis	59712768	M	40H
Vi.fis	59712843	M	40H
Vi.fis	59713275	M	40H
Vi.fis	59713290	M	40H
Vi.fis	59713352	MM	40H
Vi.fis	59713353	MM	40H
Vi.fis	59713429	M	40H

Table A.13 (continued)

ID	GI	Gene Neighborhood	Class
Vi.fis	59713483	M-M	40H
Vi.fis	59713485	M-M	40H
Vi.fis	59713508	M	40H
Vi.fis	59713572	M	40H
Vi.fis	59713630	MM	40H
Vi.fis	59713631	MM	40H
Vi.fis	59713664	M	40H
Vi.fis	59713710	MM	40H
Vi.fis	59713711	MM	40H
Vi.fis	59713860	M	40H
Vi.fis	59714042	M-----M	40H
Vi.fis	59714048	M-----M	40H
Vi.fis	59714074	M	40H
Vi.fis	59714252	mmM-M	40H
Vi.fis	59714254	mmM-M	40H
Vi.fis	59714255	m-mMM	40H
Vi.fis	59714256	m-mMM	40H
Vi.fis	59714267	M	40H
Vi.par	28896957	M	40H
Vi.par	28897196	M	40H
Vi.par	28897737	M	40H
Vi.par	28897862	M	40H
Vi.par	28897959	M	40H
Vi.par	28898260	C1---M	40H
Vi.par	28898402	M	40H
Vi.par	28898666	M	40H
Vi.par	28898678	M	40H
Vi.par	28898755	M	40H
Vi.par	28898933	M	40H
Vi.par	28899403	M	40H
Vi.par	28899601	M	40H
Vi.par	28899879	M	40H
Vi.par	28900054	M	40H
Vi.par	28900346	M	40H
Vi.par	28900366	M	40H
Vi.par	28900409	M	24H
Vi.par	28900417	M	40H
Vi.par	28900451	M	40H
Vi.par	28900467	M	24H
Vi.par	28900697	M	40H
Vi.par	28900855	M	40H
Vi.par	28901037	M	40H
Vi.par	28901044	M	40H
Vi.par	28901304	M	40H
Vi.par	28901317	M	40H
Vi.par	28901347	M	40H

Table A.13 (continued)

ID	GI	Gene Neighborhood	Class
Vi.par	28901506	M	40H
Vi.par	28901548	M	
Vi.vul	27363527	M	40H
Vi.vul	27363668	M	40H
Vi.vul	27364074	M	40H
Vi.vul	27364275	M	40H
Vi.vul	27364632	M	40H
Vi.vul	27364882	M	40H
Vi.vul	27365130	M	40H
Vi.vul	27365157	M	40H
Vi.vul	27365391	M	40H
Vi.vul	27365400	M	40H
Vi.vul	27365429	M	40H
Vi.vul	27365442	M	40H
Vi.vul	27365488	M	
Vi.vul	27365528	M	40H
Vi.vul	27365568	M	40H
Vi.vul	27365579	M	40H
Vi.vul	27365673	M	40H
Vi.vul	27365863	C1---M	40H
Vi.vul	27365879	M	40H
Vi.vul	27366002	M	40H
Vi.vul	27366013	M	40H
Vi.vul	27366036	M	40H
Vi.vul	27366148	M	40H
Vi.vul	27366407	M	40H
Vi.vul	27366502	M	40H
Vi.vul	27366531	M---v3	40H
Vi.vul	27366677	M	40H
Vi.vul	27366778	M	40H
Vi.vul	27366798	M	40H
Vi.vul	27366812	M----R2B2	40H
Vi.vul	27366840	MM	40H
Vi.vul	27366841	MM	40H
Vi.vul	27366911	M	40H
Vi.vul	27366961	M	40H
Vi.vul	27366987	M	40H
Vi.vul	27367232	M	40H
Vi.vul	27367374	M	40H
Vi.vul	27367433	M	40H
Vi.vul	27367488	M	40H
Vi.vul	27367542	Y2A2W4W3MR3DB3M	
Vi.vul	27367546	Y2A2W4W3MR3DB3M	36H
Vi.vul	27367575	M	40H
Vi.vul	27367630	M	40H
Vi.vul	27367639	M	40H

Table A.13 (continued)

ID	GI	Gene Neighborhood	Class
Vi.vul	27367871	M	40H
Vi.vul	27367899	M-M	40H
Vi.vul	27367901	M-M	40H
Vi.vul	27367913	M	40H
Vi.vul	27367965	MM---V4	40H
Vi.vul	27367966	MM---V4	40H
Vi.vul	27367986	M	40H
Vi.vul	27367999	M	
Wo.suc	34556547	W1M	36H
Wo.suc	34556577	M	
Wo.suc	34556720	M	28H
Wo.suc	34556859	M	28H
Wo.suc	34557060	M	40H
Wo.suc	34557108	M	
Wo.suc	34557162	M	40H
Wo.suc	34557239	M	28H
Wo.suc	34557253	M----M	40H
Wo.suc	34557258	M----M	28H
Wo.suc	34557318	M	
Wo.suc	34557328	M	28H
Wo.suc	34557338	M	
Wo.suc	34557389	M	40H
Wo.suc	34557432	M	
Wo.suc	34557576	M--br--M	28H
Wo.suc	34557583	M--br--M	28H
Wo.suc	34557595	M	28H
Wo.suc	34557692	M--m	28H
Wo.suc	34557695	M--m	
Wo.suc	34557756	M	40H
Wo.suc	34557844	M	
Wo.suc	34557853	M	
Wo.suc	34557985	M	
Wo.suc	34558144	M	28H
Wo.suc	34558210	M	28H
Wo.suc	34558217	M	40H
Wo.suc	34558241	M----m-M	28H
Wo.suc	34558246	m-M----m	
Wo.suc	34558248	M----m-M	28H
Wo.suc	34558397	M	40H
Wo.suc	34558440	M	28H
Xa.axo	21241382	M	36H
Xa.axo	21242415	M	36H
Xa.axo	21242494	M	36H
Xa.axo	21242636	W5-Y2A1M-MM-MMMMMMMR2DB2	36H
Xa.axo	21242637	W5-Y2A1M-MM-MMMMMMMR2DB2	36H
Xa.axo	21242638	W5-Y2A1M-MM-MMMMMMMR2DB2	36H

Table A.13 (continued)

ID	GI	Gene Neighborhood	Class
Xa.axo	21242640	W5-Y2A1M-MM-MMMMMMMR2DB2	36H
Xa.axo	21242641	W5-Y2A1M-MM-MMMMMMMR2DB2	36H
Xa.axo	21242646	W5-Y2A1M-MM-MMMMMMMR2DB2	36H
Xa.axo	21242740	M	
Xa.axo	21243181	MW1	34H
Xa.axo	21243593	A2MW2-R3B3	34H
Xa.axo	21243826	Y6Y5W4MA3B4W3	40H
Xa.axo	21243939	M	36H
Xa.axo	21244174	M	
Xa.axo	21244493	M	36H
Xa.axo	77748616	W5-Y2A1M-MM-MMMMMMMR2DB2	36H
Xa.axo	77748617	W5-Y2A1M-MM-MMMMMMMR2DB2	36H
Xa.axo	77748618	W5-Y2A1M-MM-MMMMMMMR2DB2	36H
Xa.axo	77748619	W5-Y2A1M-MM-MMMMMMMR2DB2	36H
Xa.axo	77748700	M	36H
Xa.cam	21229754	M	36H
Xa.cam	21229802	M	36H
Xa.cam	21231179	M	36H
Xa.cam	21231317	W6-Y2A1M--MMMM-MM-M-W1-MR2DB2	36H
Xa.cam	21231321	W6-Y2A1M--MMMM-MM-M-W1-MR2DB2	36H
Xa.cam	21231327	W6-Y2A1M--MMMM-MM-M-W1-MR2DB2	36H
Xa.cam	21231329	W6-Y2A1M--MMMM-MM-M-W1-MR2DB2	36H
Xa.cam	21231332	W6-Y2A1M--MMMM-MM-M-W1-MR2DB2	36H
Xa.cam	21231410	M	
Xa.cam	21231495	M	36H
Xa.cam	21231752	MW2	34H
Xa.cam	21232132	A3MW3-R3B4	34H
Xa.cam	21232353	Y6Y5W5MA4B5W4	40H
Xa.cam	21232514	M	36H
Xa.cam	21232750	M	
Xa.cam	21232951	M	36H
Xa.cam	77747851	W6-Y2A1M--MMMM-MM-M-W1-MR2DB2	36H
Xa.cam	77747852	W6-Y2A1M--MMMM-MM-M-W1-MR2DB2	36H
Xa.cam	77747853	W6-Y2A1M--MMMM-MM-M-W1-MR2DB2	36H
Xa.cam	77747854	W6-Y2A1M--MMMM-MM-M-W1-MR2DB2	36H
Xa.ory	58581092	A1MW1-R1B1	34H
Xa.ory	58581372	Y6Y2W2MA2A2B2W3	40H
Xa.ory	58582181	M	
Xa.ory	58582461	W4-Y5A4-MM-M-MMMM-W5----R2DB3	36H
Xa.ory	58582463	W4-Y5A4-MM-M-MMMM-W5----R2DB3	36H
Xa.ory	58582467	W4-Y5A4-MM-M-MMMM-W5----R2DB3	36H
Xa.ory	58582468	W4-Y5A4-MM-M-MMMM-W5----R2DB3	36H
Xa.ory	58582470	W4-Y5A4-MM-M-MMMM-W5----R2DB3	36H
Xa.ory	77760738	W4-Y5A4-MM-M-MMMM-W5----R2DB3	36H
Xa.ory	77760739	W4-Y5A4-MM-M-MMMM-W5----R2DB3	36H
Xa.ory	15838547	Y2W2MABW1	40H

Table A.13 (continued)

ID	GI	Gene Neighborhood	Class
Ye.pes	16121114	M	36H
Ye.pes	16121516	M	36H
Ye.pes	16121912	M	36H
Ye.pes	16121939	MRBYZ	36H
Ye.pes	16122736	M	24H
Ye.pes	16124176	M	36H
Ye.pse	51595609	M	36H
Ye.pse	51596723	m-----AW--MMRBYZ	36H
Ye.pse	51596724	m-----AW--MMRBYZ	36H
Ye.pse	51596734	zybrmm--wa-----M	36H
Ye.pse	51596746	M	36H
Ye.pse	51596870	M	24H
Ye.pse	51597365	M	36H
Ye.pse	51598209	M	36H
Zy.mob	56550981	M-A1RB1DYW	36H
Zy.mob	56551098	M	36H
Zy.mob	56551778	M-A2-B2	36H

Table A.14 FlhA data. ID, GI, and Range are explained in Table A.2. One asterisk (*) marks sequences that were not included in the phylogenetic analysis due to significant deletions. The FlhA from *Bacillus cereus* ATCC 10987 was used because FlhA was not annotated in the proteome of *Bacillus cereus* ATCC 14579 (although it is in an unidentified ORF based on BLAST searches against the genome). A minimum evolution tree was built from the FlhA alignment in MEGA with pairwise deletions and the JTT distance matrix.

ID	GI	Range
Ac.bac	94968666	18-691
Ag.tum	15887926	47-722
An.deh	86157784	17-684
Aq.aeo	15606449	6-675
Azoar*	56475795	19-201
Ba.ant	49184608	15-692
Ba.cer	42780862	10-687
Ba.cla	56964015	6-675
Ba.hal	15615001	6-674
Ba.lic	52080242	6-675
Ba.sub	16078702	6-675
Ba.thu	49477323	15-692
Bd.bac	42524690	20-693
Bo.bro	33601533	19-700
Bo.bur	15594616	21-693
Bo.gar	51598531	20-692
Br.jap1	27377318	66-747
Br.jap2	27381962	19-692
Br.mel-N*	17988510	20-99
Br.mel-C*	17988511	1-588
Br.sui	23500847	47-722
Bu.aph	21672517	28-699
Bu.cen	107024382	24-696
Bu.mal	53724311	23-696
Bu.pse	53720904	23-696
Bu.tha1	83716612	16-695
Bu.tha2	83719420	23-696
Bu.xen	91785646	23-696
Burkh	78064848	23-695
Ca.cre	16125162	27-700
Ca.hyd	78044126	17-682
Ca.jej	15792213	27-720
Ch.sal	92114137	22-691
Ch.vio1	34496480	16-689
Ch.vio2	34498462	1-569
Cl.ace	15895416	17-688
Cl.tet	28211316	1-607
Co.psy	71282152	21-698
De.aro	71906379	20-690
De.des	78355427	21-700

Table A.14 (continued)

ID	GI	Range
De.haf	89895720	15-682
De.psy	51246527	24-697
De.vul	46581634	22-700
Er.car	50120631	22-689
Es.col1	15799934	1-578
Es.col2	15802291	22-689
Ge.kau	56419773	8-678
Ge.met	78221647	20-689
Ge.sul	39998147	20-691
Gl.oxy	58038895	37-710
Ha.che1	83646790	17-679
Ha.che2	83647842	18-720
He.aci	109947677	26-729
He.hep	32265966	25-731
He.pyl	15611451	26-729
Id.loi	56460228	25-703
Janna	89056677	12-691
La.int	94986972	18-693
Le.int	24215307	11-691
Le.pne	52842012	16-691
Li.inn	16799763	10-688
Li.mon	16802722	10-688
Ma.mag	83309597	28-703
Me.fla	91776297	16-686
Me.lot	13472610	19-694
Mo.the	83589642	20-687
Ni.eur	30250407	19-689
Ni.ham	92118828	66-745
Ni.mul	82702439	19-688
Ni.oce	77165623	18-688
Ni.win	75674727	32-709
Oc.ihe	23099030	6-674
Pe.car	77918766	19-689
Ph.lum	37525819	22-689
Ph.pro1	54307249	17-699
Ph.pro2	54308129	22-696
Ps.aer	15596649	22-704
Ps.atl	109899338	22-698
Ps.ent	104782804	22-705
Ps.flu	70729054	22-705
Ps.hal	77359755	1-675
Ps.put	26991034	22-705
Ps.syr	28869180	22-705
Ra.eut	73539438	22-691
Ra.met	94312631	22-691
Ra.sol	17549612	20-692

Table A.14 (continued)

ID	GI	Range
Rh.bal	32475306	10-685
Rh.etl	86356337	19-694
Rh.fer	89902467	22-688
Rh.pal	39934708	29-710
Rh.rub	83591878	45-720
Rh.sph1	77463600	17-692
Rh.sph2	77464897	13-686
Sa.deg	90021812	33-744
Sa.ent	16760862	22-689
Sa.rub	83816721	1-672
Sh.boy	82543641	22-689
Sh.den1	91791413	14-672
Sh.den2	91792698	21-697
Sh.den3	91795015	15-692
Sh.fle	30061846	1-578
Sh.one	24374725	21-697
Sh.son	74310873	1-578
Si.mel	15964439	19-694
Si.pom	56695093	14-693
Silic	99082793	14-694
So.glo	85058004	21-693
Sp.ala	103488380	17-685
Sy.aci	85859184	20-689
Sy.the	51894123	15-681
Tb.den1	74317195	16-686
Tb.den2	74317264	19-688
Th.mar	15643670	6-676
Th.ten	20807869	6-668
Tm.cru	78485089	18-693
Tm.den	78776923	24-719
Tr.den	42525574	13-693
Tr.pal	15639701	13-693
Vi.cho	15642069	22-695
Vi.fis	59712444	22-695
Vi.par1	28899009	22-697
Vi.par2	28901401	16-695
Vi.vul	27365294	24-698
Wi.glo	32490784	41-719
Wo.suc	34557437	26-745
Xa.axo	21242680	21-696
Xa.cam	21231357	21-696
Xa.ory	58582241	21-696
Ye.pes	16122045	22-689
Ye.pse1	51596002	22-689
Ye.pse2	51597657	18-697
Zy.mob	56551520	18-689

Table A.15 FlaH data. ID, GI, and Range are explained in Table A.2. A minimum evolution tree was built from the FlaH alignment in MEGA with pairwise deletions and the Poisson correction distance.

ID	GI	Range
Ae.per	14601707	20-243
Ar.ful	11498655	21-242
Ha.mar	55378882	16-253
Halob	16554479	14-256
Me.ace1	20091875	15-236
Me.ace2	20091899	25-246
Me.bar	73669467	39-260
Me.bur1	91772404	15-236
Me.bur2	91773536	11-231
Me.hun	88601422	33-254
Me.jan	15669089	13-233
Me.mar	45359237	10-230
Me.maz1	21226420	15-236
Me.maz2	21226516	41-262
Na.pha	76801722	15-253
Py.aby	14521687	12-232
Py.fur	18976704	1-204
Py.hor	14590456	12-232
Su.aci	70606944	7-226
Su.sol	15899077	12-232
Su.tok	15922857	12-232
Th.aci	16081662	9-223
Th.kod	57639982	12-232
Th.vol	13541443	9-223

Table A.16 PilT and PilU data. ID, GI, and Range are explained in Table A.2. Sequences are classified as PilT or PilU proteins after differentiation in phylogenetic analysis. An asterisk (*) marks sequences not included in the final analysis due to a deletrious frameshift. A minimum evolution tree was built from the alignment of PilT and PilU seqeunces in MEGA with pairwise deletions and the Poisson correction distance.

ID	GI	Class	Range
Ac.bac1	94967720	PilU	22-356
Ac.bac2	94967735	PilU	3-333
Ac.bac3	94968419	PilT	5-333
Acine1	50084119	PilU	3-331
Acine2	50084120	PilT	3-329
An.deh1	86157051	PilT	4-332
An.deh2	86159136	PilU	12-342
An.deh3	86159787	PilU	3-333
An.var1	75906600	PilT	5-336
An.var2	75909865	PilU	83-408
Aq.aeo1	15606134	PilT	15-343
Aq.aeo2	15607093	PilU	2-334
Azoar1	56476154	PilU	14-341
Azoar2	56476395	PilT	3-330
Azoar3	56476396	PilU	10-337
Azoar4	56479088	PilU	3-329
Bd.bac	42525175	PilT	4-331
Bo.bro	33599780	PilT	3-329
Ca.hyd	78044094	PilT	2-342
Ca.Pel	71082782	PilT	5-329
Ch.sal1	92115167	PilU	7-334
Ch.sal2	92115168	PilT	3-329
Ch.vio1	34495634	PilT	3-330
Ch.vio2	34495635	PilU	10-337
Ch.vio3	34496913	PilU	3-329
Cl.ace	15894967	PilT	4-331
Cl.per	18310749	PilT	4-329
Cl.tet	28210816	PilT	4-330
Co.psy1	71281718	PilT	3-329
Co.psy2	71282483	PilU	3-330
De.aro1	71907355	PilU	3-329
De.aro2	71909493	PilU	10-337
De.aro3	71909494	PilT	3-330
De.eth1	57233890	PilT	3-331
De.eth2	57233979	PilT	4-332
De.eth3	57234308	PilT	3-331
De.geo1	94984364	PilU	4-334
De.geo2	94984372	PilT	8-335
De.geo3	94984776	PilU	3-332
De.haf	89895131	PilT	7-334
De.psy1	51244748	PilU	6-342

Table A.16 (continued)

ID	GI	Class	Range
De.psy2	51245546	PilT	4-331
De.rad1	15805469	PilU	4-334
De.rad2	15806182	PilU	3-332
De.rad3	15806961	PilT	9-336
De.vul1	46579673	PilT	4-331
De.vul2	46579681	PilT	5-330
Dehal1	73748690	PilT	3-331
Dehal2	73748939	PilT	4-332
Dehal3	73749044	PilT	3-331
Er.car	50122547	PilT	3-330
Fr.tul	56707266	PilT	2-330
Ge.kau	56421163	PilT	5-332
Ge.met1	78221423	PilU	3-333
Ge.met2	78221483	PilU	6-334
Ge.met3	78222606	PilT	4-332
Ge.met4	78224590	PilT	4-331
Ge.sul1	39995257	PilU	3-333
Ge.sul2	39995340	PilU	6-334
Ge.sul3	39995544	PilT	4-331
Ge.sul4	39996592	PilT	4-332
Gl.vio	37522235	PilT	6-336
Ha.che1	83647692	PilU	3-329
Ha.che2	83648987	PilU	3-330
Ha.che3	83648988	PilT	3-329
Id.loi	56461073	PilT	3-329
Le.pne1	52842230	PilT	3-329
Le.pne2	52842533	PilU	3-329
Me.cap1	53804163	PilU	3-330
Me.cap2	53804164	PilT	3-329
Me.fla1	91775464	PilU	10-336
Me.fla2	91775740	PilU	6-332
Me.fla3	91776458	PilU	10-337
Me.fla4	91776459	PilT	3-330
Mo.the	83590396	PilT	4-329
My.xan1	108756799	PilT	4-332
My.xan2	108761365	PilU	13-342
My.xan3	108762281	PilT	195-521
My.xan4	108763066	PilT	6-333
My.xan5	108763886	PilU	26-356
Ne.gon1	59800790	PilU	7-334
Ne.gon2	59802222	PilT	3-330
Ne.gon3	59802223	PilU	39-366
Ne.men1	15675989	PilU	39-366
Ne.men2	15675990	PilT	3-330
Ne.men3	15676666	PilU	7-334

Table A.16 (continued)

ID	GI	Class	Range
Ni.eur1	30248967	PilU	10-337
Ni.eur2	30248968	PilT	3-330
Ni.oce1	77164860	PilU	3-329
Ni.oce2	77166451	PilT	3-329
Nosto1	17229935	PilT	5-336
Nosto2	17230821	PilU	84-409
Pe.car1	77918120	PilT	4-331
Pe.car2	77918481	PilT	17-344
Pe.car3	77919742	PilT	12-340
Pe.car4	77920000	PilU	3-333
Ph.lum	37525149	PilT	3-331
Ph.pro1	54310238	PilU	5-330
Ph.pro2	54310239	PilT	3-329
Polar1	91786305	PilU	10-337
Polar2	91786306	PilT	3-330
Polar3	91787871	PilU	10-336
Pr.mar1	33862292	PilT	5-336
Pr.mar2	33863581	PilU	75-400
Ps.aer1	15595592	PilT	3-329
Ps.aer2	15595593	PilU	3-330
Ps.arc1	71064843	PilU	3-331
Ps.arc2	71064844	PilT	8-334
Ps.atl1	109900017	PilU	3-330
Ps.atl2	109900018	PilT	3-329
Ps.cry1	93005123	PilU	3-331
Ps.cry2	93005124	PilT	8-334
Ps.ent	104779599	PilT	3-329
Ps.flu	70733122	PilT	74-400
Ps.hal1	77361517	PilU	5-331
Ps.hal2	77361518	PilT	1-303
Ps.put	26991769	PilT	7-333
Ps.syr1	28870209	PilU	5-331
Ps.syr2	28872159	PilT	3-329
Ra.eut1	73542484	PilU	11-338
Ra.eut2	73542485	PilT	3-330
Ra.met1	94311867	PilU	13-340
Ra.met2	94311868	PilT	3-330
Ra.sol1	17547400	PilU	11-338
Ra.sol2	17547401	PilT	3-330
Rh.bal1	32476325	PilU	39-370
Rh.bal2	32477702	PilT	6-332
Rh.fer1	89900830	PilU	16-342
Rh.fer2	89902636	PilU	10-337
Rh.fer3	89902637	PilT	3-330
Rh.rub	83591458	PilT	3-339
Ru.xyl1	108804006	PilT	10-337

Table A.16 (continued)

ID	GI	Class	Range
Ru.xyl2	108804008	PilU	21-349
Sa.deg1	90021009	PilU	5-331
Sa.deg2	90023279	PilU	3-330
Sa.deg3	90023280	PilT	3-329
Sa.ent	16761874	PilT	3-326
Sh.boy	82545426	PilT	18-341
Sh.den1	91794031	PilU	3-330
Sh.den2	91794032	PilT	7-333
Sh.dys	82778277	PilT	18-341
Sh.fle	30064266	PilT	18-341
Sh.one1	24374861	PilU	3-330
Sh.one2	24374862	PilT	3-329
Sh.son	74313508	PilT	18-341
Sy.aci1	85858696	PilT	4-331
Sy.aci2	85858697	PilU	20-356
Sy.elo1	56750702	PilU	6-333
Sy.elo2	56751763	PilU	49-374
Sy.elo3	56752032	PilT	5-336
Synco1	86606611	PilT	2-333
Synco2	86607393	PilU	106-432
Syncy1	16331035	PilU	80-405
Syncy2	16331158	PilT	7-338
SyneC	78185291	PilU	40-356
SyneW-N*	33866363	PilU	1-152
SyneW-C*	33866364	PilU	50-212
Tb.den1	74318437	PilT	3-330
Tb.den2	74318438	PilU	9-336
Th.elo1	22297665	PilT	5-336
Th.elo2	22298097	PilU	97-423
Th.mar	15644114	PilT	8-336
Th.ten	20807712	PilT	4-330
Th.the1	46199717	PilU	12-341
Th.the2	46199923	PilT	7-334
Tm.cru	78486091	PilT	3-329
Tm.den	78777449	PilT	11-338
Vi.cho1	15640489	PilT	3-329
Vi.cho2	15640490	PilU	3-328
Vi.fis1	59711038	PilT	3-329
Vi.fis2	59711039	PilU	3-328
Vi.par1	28899388	PilU	3-328
Vi.par2	28899389	PilT	3-329
Vi.vul1	27364897	PilT	3-329
Vi.vul2	27364898	PilU	3-328
Wo.suc1	34556951	PilU	22-348
Wo.suc2	34556954	PilT	6-333
Xa.axo1	21243650	PilU	6-333

Table A.16 (continued)

ID	GI	Class	Range
Xa.axo2	21243651	PilT	3-329
Xa.axo3	21244332	PilU	3-329
Xa.cam1	21230068	PilU	5-331
Xa.cam2	21232185	PilU	6-333
Xa.cam3	21232186	PilT	3-329
Xa.ory1	58580396	PilU	5-331
Xa.ory2	58581041	PilT	3-329
Xa.ory3	58581042	PilU	23-350
Xy.fas1	15838233	PilU	6-333
Xy.fas2	15838234	PilT	3-329
Ye.pes	16121244	PilT	27-354
Ye.pse	51597522	PilT	43-370

REFERENCES

1. Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.F., Dougherty, B.A., Merrick, J.M. *et al.* (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science*, **269**, 496-512.

2. Leite, M. (2004) Public sphere and the sustainability of the bioinformatics promise. *Genet Mol Res*, **3**, 575-581.

3. Glaser, P., Frangeul, L., Buchrieser, C., Rusniok, C., Amend, A., Baquero, F., Berche, P., Bloecker, H., Brandt, P., Chakraborty, T. *et al.* (2001) Comparative genomics of Listeria species. *Science*, **294**, 849-852.

4. Han, C.S., Xie, G., Challacombe, J.F., Altherr, M.R., Bhotika, S.S., Brown, N., Bruce, D., Campbell, C.S., Campbell, M.L., Chen, J. *et al.* (2006) Pathogenomic sequence analysis of Bacillus cereus and Bacillus thuringiensis isolates closely related to Bacillus anthracis. *J Bacteriol*, **188**, 3382-3390.

5. Snyder, L.A. and Saunders, N.J. (2006) The majority of genes in the pathogenic Neisseria species are present in non-pathogenic Neisseria lactamica, including those designated as 'virulence genes'. *BMC genomics*, **7**, 128.

6. Jordan, I.K., Kondrashov, F.A., Adzhubei, I.A., Wolf, Y.I., Koonin, E.V., Kondrashov, A.S. and Sunyaev, S. (2005) A universal trend of amino acid gain and loss in protein evolution. *Nature*, **433**, 633-638.

7. Brandvain, Y., Barker, M.S. and Wade, M.J. (2007) Gene co-inheritance and gene transfer. *Science*, **315**, 1685.

8. Makarova, K., Slesarev, A., Wolf, Y., Sorokin, A., Mirkin, B., Koonin, E., Pavlov, A., Pavlova, N., Karamychev, V., Polouchine, N. *et al.* (2006) Comparative genomics of the lactic acid bacteria. *Proc Natl Acad Sci U S A*, **103**, 15611-15616.

9. Koonin, E.V. and Wolf, Y.I. (2006) Evolutionary systems biology: links between gene evolution and function. *Current opinion in biotechnology*, **17**, 481-487.

10. Ulrich, L.E., Koonin, E.V. and Zhulin, I.B. (2005) One-component systems dominate signal transduction in prokaryotes. *Trends in Microbiology*, **13**, 52-56.

11. Parkinson, J.S. and Kofoid, E.C. (1992) Communication modules in bacterial signaling proteins. *Annu Rev Genet*, **26**, 71-112.

12. Parkinson, J.S. (1993) Signal transduction schemes of bacteria. *Cell*, **73**, 857-871.

13. Hoch, J.A. (2000) Two-component and phosphorelay signal transduction. *Curr Opin Microbiol*, **3**, 165-170.

14. Stock, A.M., Robinson, V.L. and Goudreau, P.N. (2000) Two-component signal transduction. *Annu Rev Biochem*, **69**, 183-215.

15. Wadhams, G.H. and Armitage, J.P. (2004) Making sense of it all: Bacterial chemotaxis. *Nat Rev Mol Cell Bio*, **5**, 1024-1037.

16. Bourret, R.B. and Stock, A.M. (2002) Molecular information processing: Lessons from bacterial chemotaxis. *Journal of Biological Chemistry*, **277**, 9625-9628.

17. Baker, M.D., Wolanin, P.M. and Stock, J.B. (2006) Signal transduction in bacterial chemotaxis. *Bioessays*, **28**, 9-22.

18. Djordjevic, S., Goudreau, P.N., Xu, Q.P., Stock, A.M. and West, A.H. (1998) Structural basis for methylesterase CheB regulation by a phosphorylation-activated domain. *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 1381-1386.

19. Djordjevic, S. and Stock, A.M. (1998) Chemotaxis receptor recognition by protein methyltransferase CheR. *Nature Structural Biology*, **5**, 446-450.

20. McEvoy, M.M., Hausrath, A.C., Randolph, G.B., Remington, S.J. and Dahlquist, F.W. (1998) Two binding modes reveal flexibility in kinase/response regulator interactions in the bacterial chemotaxis pathway. *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 7333-7338.

21. Welch, M., Chinardet, N., Mourey, L., Birck, C. and Samama, J.P. (1998) Structure of the CheY-binding domain of histidine kinase CheA in complex with CheY. *Nature Structural Biology*, **5**, 25-29.

22. Bilwes, A.M., Alex, L.A., Crane, B.R. and Simon, M.I. (1999) Structure of CheA, a signal-transducing histidine kinase. *Cell*, **96**, 131-141.

23. Kim, K.K., Yokota, H. and Kim, S.H. (1999) Four-helical-bundle structure of the cytoplasmic domain of a serine chemotaxis receptor. *Nature*, **400**, 787-792.

24. Griswold, I.J., Zhou, H.J., Matison, M., Swanson, R.V., McIntosh, L.P., Simon, M.I. and Dahlquist, F.W. (2002) The solution structure and interactions of CheW from Thermotoga maritima. *Nature Structural Biology*, **9**, 121-125.

25. Zhao, R., Collins, E.J., Bourret, R.B. and Silversmith, R.E. (2002) Structure and catalytic mechanism of the E. coli chemotaxis phosphatase CheZ. *Nature Structural Biology*, **9**, 570-575.

26. Park, S.Y., Beel, B.D., Simon, M.I., Bilwes, A.M. and Crane, B.R. (2004) In different organisms, the mode of interaction between two signaling proteins is not necessarily conserved. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 11646-11651.

27. Park, S.Y., Chao, X.J., Gonzalez-Bonet, G., Beel, B.D., Bilwes, A.M. and Crane, B.R. (2004) Structure and function of an unusual family of protein phosphatases: The bacterial chemotaxis proteins CheC and CheX. *Mol Cell*, **16**, 563-574.

28. Quezada, C.M., Gradinaru, C., Simon, M.I., Bilwes, A.M. and Crane, B.R. (2004) Helical shifts generate two distinct conformers in the atomic resolution structure of the CheA phosphotransferase domain from Thermotoga maritima. *Journal of Molecular Biology*, **341**, 1283-1294.

29. Chao, X., Muff, T.J., Park, S.-Y., Zhang, S., Pollard, A.M., Ordal, G.W., Bilwes, A.M. and Crane, B.R. (2006) A Receptor-Modifying Deamidase in Complex with a Signaling Phosphatase Reveals Reciprocal Regulation. *Cell*, **124**, 561-571.

30. Guhaniyogi, J., Robinson, V.L. and Stock, A.M. (2006) Crystal structures of beryllium fluoride-free and beryllium fluoride-bound CheY in complex with the conserved C-terminal peptide of CheZ reveal dual binding modes specific to CheY conformation. *J Mol Biol*, **359**, 624-645.

31. Park, S.Y., Borbat, P.P., Gonzalez-Bonet, G., Bhatnagar, J., Pollard, A.M., Freed, J.H., Bilwes, A.M. and Crane, B.R. (2006) Reconstruction of the chemotaxis receptor-kinase assembly. *Nature Structural & Molecular Biology*, **13**, 400-407.

32. Mourey, L., Da Re, S., Pedelacq, J.D., Tolstykh, T., Faurie, C., Guillet, V., Stock, J.B. and Samama, J.P. (2001) Crystal structure of the CheA histidine phosphotransfer domain that mediates response regulator phosphorylation in bacterial chemotaxis. *J Biol Chem*, **276**, 31074-31082.

33. Szurmant, L. and Ordal, G.W. (2004) Diversity in chemotaxis mechanisms among the bacteria and archaea. *Microbiology and Molecular Biology Reviews*, **68**, 301-319.

34. Alexander, R.P. and Zhulin, I.B. (2007) Evolutionary genomics reveals conserved structural determinants of signaling and adaptation in microbial chemoreceptors. *Proc Natl Acad Sci U S A*, **104**, 2885-2890.

35. Kristich, C.J. and Ordal, G.W. (2002) Bacillus subtilis CheD is a chemoreceptor modification enzyme required for chemotaxis. *Journal of Biological Chemistry*, **277**, 25356-25362.

36. Rosario, M.M.L. and Ordal, G.W. (1996) CheC and CheD interact to regulate methylation of Bacillus subtilis methyl-accepting chemotaxis proteins. *Molecular Microbiology*, **21**, 511-518.

37. Karatan, E., Saulmon, M.M., Bunn, M.W. and Ordal, G.W. (2001) Phosphorylation of the response regulator CheV is required for adaptation to attractants during Bacillus subtilis chemotaxis. *Journal of Biological Chemistry*, **276**, 43618-43626.

38. Francis, N.R., Wolanin, P.M., Stock, J.B., DeRosier, D.J. and Thomas, D.R. (2004) Three-dimensional structure and organization of a receptor/signaling complex. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 17480-17485.

39. Maddock, J.R. and Shapiro, L. (1993) Polar Location of the Chemoreceptor Complex in the Escherichia-Coli Cell. *Science*, **259**, 1717-1723.

40. Martin, A.C., Wadhams, G.H. and Armitage, J.P. (2001) The roles of the multiple CheW and CheA homologues in chemotaxis and in chemoreceptor localization in Rhodobacter sphaeroides. *Mol Microbiol*, **40**, 1261-1272.

41. Guvener, Z.T., Tifrea, D.F. and Harwood, C.S. (2006) Two different Pseudomonas aeruginosa chemosensory signal transduction complexes localize to cell poles and form and remould in stationary phase. *Mol Microbiol*, **61**, 106-118.

42. Thompson, S.R., Wadhams, G.H. and Armitage, J.P. (2006) The positioning of cytoplasmic protein clusters in bacteria. *PNAS*, 0600919103.

43. Wadhams, G.H., Martin, A.C., Porter, S.L., Maddock, J.R., Mantotta, J.C., King, H.M. and Armitage, J.P. (2002) TlpC, a novel chemotaxis protein in Rhodobacter sphaeroides, localizes to a discrete region in the cytoplasm. *Molecular Microbiology*, **46**, 1211-1221.

44. Eddy, S.R. (1996) Hidden Markov models. *Curr Opin Struc Biol*, **6**, 361-365.

45. Krogh, A., Brown, M., Mian, I.S., Sjolander, K. and Haussler, D. (1994) Hidden Markov models in computational biology. Applications to protein modeling. *J Mol Biol*, **235**, 1501-1531.

46. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, **25**, 3389-3402.

47. Bhaya, D. (2004) Light matters: phototaxis and signal transduction in unicellular cyanobacteria. *Molecular Microbiology*, **53**, 745-754.

48. Porter, S.L., Warren, A.V., Martin, A.C. and Armitage, J.P. (2002) The third chemotaxis locus of Rhodobacter sphaeroides is essential for chemotaxis. *Mol Microbiol*, **46**, 1081-1094.

49. Vlamakis, H.C., Kirby, J.R. and Zusman, D.R. (2004) The Che4 pathway of Myxococcus xanthus regulates type IV pilus-mediated motility. *Mol Microbiol*, **52**, 1799-1811.

50. Whitchurch, C.B., Leech, A.J., Young, M.D., Kennedy, D., Sargent, J.L., Bertrand, J.J., Semmler, A.B., Mellick, A.S., Martin, P.R., Alm, R.A. *et al.* (2004) Characterization of a complex chemosensory signal transduction system which controls twitching motility in Pseudomonas aeruginosa. *Mol Microbiol*, **52**, 873-893.

51. Berleman, J.E. and Bauer, C.E. (2005) A che-like signal transduction cascade involved in controlling flagella biosynthesis in Rhodospirillum centenum. *Molecular Microbiology*, **55**, 1390-1402.

52. Hickman, J.W., Tifrea, D.F. and Harwood, C.S. (2005) A chemosensory system that regulates biofilm formation through modulation of cyclic diguanylate levels. *Proc Natl Acad Sci U S A*, **102**, 14422-14427.

53. Berleman, J.E. and Bauer, C.E. (2005) Involvement of a Che-like signal transduction cascade in regulating cyst cell development in Rhodospirillum centenum. *Mol Microbiol*, **56**, 1457-1466.

54. Kirby, J.R. and Zusman, D.R. (2003) Chemosensory regulation of developmental gene expression in Myxococcus xanthus. *Proc Natl Acad Sci U S A*, **100**, 2008-2013.

55. Rosario, M.M.L., Kirby, J.R., Bochar, D.A. and Ordal, G.W. (1995) Chemotactic Methylation and Behavior in Bacillus-Subtilis - Role of 2 Unique Proteins, Chec and Ched. *Biochemistry*, **34**, 3823-3831.

56. Sourjik, V., Sterr, W., Platzer, J., Bos, I., Haslbeck, M. and Schmitt, R. (1998) Mapping of 41 chemotaxis, flagellar and motility genes to a single region of the Sinorhizobium meliloti chromosome. *Gene*, **223**, 283-290.

57. Gosink, K.K., Kobayashi, R., Kawagishi, I. and Hase, C.C. (2002) Analyses of the roles of the three cheA homologs in chemotaxis of Vibrio cholerae. *Journal of Bacteriology*, **184**, 1767-1771.

58. McBride, M.J., Weinberg, R.A. and Zusman, D.R. (1989) "Frizzy" aggregation genes of the gliding bacterium Myxococcus xanthus show sequence similarities to the chemotaxis genes of enteric bacteria. *Proc Natl Acad Sci U S A*, **86**, 424-428.

59. Yang, Z., Geng, Y., Xu, D., Kaplan, H.B. and Shi, W. (1998) A new set of chemotaxis homologues is essential for Myxococcus xanthus social motility. *Mol Microbiol*, **30**, 1123-1130.

60. Jiang, Z.Y. and Bauer, C.E. (1997) Analysis of a chemotaxis operon from Rhodospirillum centenum. *J Bacteriol*, **179**, 5712-5719.

61. Hauwaerts, D., Alexandre, G., Das, S.K., Vanderleyden, J. and Zhulin, I.B. (2002) A major chemotaxis gene cluster in Azospirillum brasilense and relationships between chemotaxis operons in alpha-proteobacteria. *FEMS Microbiol Lett*, **208**, 61-67.

62. Ferrandez, A., Hawkins, A.C., Summerfield, D.T. and Harwood, C.S. (2002) Cluster II che genes from Pseudomonas aeruginosa are required for an optimal chemotactic response. *J Bacteriol*, **184**, 4374-4383.

63. Bhaya, D., Takahashi, A. and Grossman, A.R. (2001) Light regulation of type IV pilus-dependent motility by chemosensor-like elements in Synechocystis PCC6803. *Proc Natl Acad Sci U S A*, **98**, 7540-7545.

64. Greene, S.R. and Stamm, L.V. (1999) Molecular characterization of a chemotaxis operon in the oral spirochete, Treponema denticola. *Gene*, **232**, 59-68.

65. Silverman, M. and Simon, M. (1977) Identification of Polypeptides Necessary for Chemotaxis in Escherichia-Coli. *Journal of Bacteriology*, **130**, 1317-1325.

66. Mayer, B.J. (2006) Protein-protein interactions in signaling cascades. *Methods in molecular biology (Clifton, N.J*, **332**, 79-99.

67. Fields, S. and Song, O. (1989) A novel genetic system to detect protein-protein interactions. *Nature*, **340**, 245-246.

68. Monti, M., Orru, S., Pagnozzi, D. and Pucci, P. (2005) Interaction proteomics. *Bioscience reports*, **25**, 45-56.

69. Fancy, D.A. (2000) Elucidation of protein-protein interactions using chemical cross-linking or label transfer techniques. *Current opinion in chemical biology*, **4**, 28-33.

70. Muronetz, V.I., Sholukh, M. and Korpela, T. (2001) Use of protein-protein interactions in affinity chromatography. *Journal of biochemical and biophysical methods*, **49**, 29-47.

71. Lueking, A., Horn, M., Eickhoff, H., Bussow, K., Lehrach, H. and Walter, G. (1999) Protein microarrays for gene expression and antibody screening. *Analytical biochemistry*, **270**, 103-111.

72. Ng, S.K., Zhang, Z. and Tan, S.H. (2003) Integrative approach for computationally inferring protein domain interactions. *Bioinformatics*, **19**, 923-929.

73. von Mering, C., Jensen, L.J., Kuhn, M., Chaffron, S., Doerks, T., Kruger, B., Snel, B. and Bork, P. (2007) STRING 7--recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res*, **35**, D358-362.

74. Goh, C.S., Bogan, A.A., Joachimiak, M., Walther, D. and Cohen, F.E. (2000) Co-evolution of proteins with their interaction partners. *Journal of Molecular Biology*, **299**, 283-293.

75. Pazos, F. and Valencia, A. (2001) Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein Engineering*, **14**, 609-614.

76. Ramani, A.K. and Marcotte, E.M. (2003) Exploiting the co-evolution of interacting proteins to discover interaction specificity. *Journal of Molecular Biology*, **327**, 273-284.

77. Snel, B., Bork, P. and Huynen, M.A. (2002) The identification of functional modules from the genomic association of genes. *Proc Natl Acad Sci U S A*, **99**, 5890-5895.

78. Valencia, A. and Pazos, F. (2002) Computational methods for the prediction of protein interactions. *Curr Opin Struc Biol*, **12**, 368-373.

79. Shukla, D., Zhu, X.Y. and Matsumura, P. (1998) Flagellar motor-switch binding face of CheY and the biochemical basis of suppression by CheY mutants that compensate for motor-switch defects in Escherichia coli. *J Biol Chem*, **273**, 23993-23999.

80. Minamino, T., Yamaguchi, S. and Macnab, R.M. (2000) Interaction between FliE and FlgB, a proximal rod component of the flagellar basal body of Salmonella. *J Bacteriol*, **182**, 3029-3036.

81. Maisnier-Patin, S., Paulander, W., Pennhag, A. and Andersson, D.I. (2007) Compensatory evolution reveals functional interactions between ribosomal proteins S12, L14 and L19. *J Mol Biol*, **366**, 207-215.

82. Moyle, W.R., Campbell, R.K., Myers, R.V., Bernard, M.P., Han, Y. and Wang, X. (1994) Co-evolution of ligand-receptor pairs. *Nature*, **368**, 251-255.

83. Gobel, U., Sander, C., Schneider, R. and Valencia, A. (1994) Correlated Mutations and Residue Contacts in Proteins. *Proteins-Structure Function and Genetics*, **18**, 309-317.

84. Suel, G.M., Lockless, S.W., Wall, M.A. and Ranganathan, R. (2003) Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nature Structural Biology*, **10**, 59-69.

85. Pazos, F., HelmerCitterich, M., Ausiello, G. and Valencia, A. (1997) Correlated mutations contain information about protein-protein interaction. *Journal of Molecular Biology*, **271**, 511-523.

86. Pazos, F. and Valencia, A. (2002) In silico two-hybrid system for the selection of physically interacting protein pairs. *Proteins-Structure Function and Genetics*, **47**, 219-227.

87. Ulrich, L.E. and Zhulin, I.B. (2007) MiST: a microbial signal transduction database. *Nucleic Acids Res*, **35**, D386-390.

88. Pruitt, K.D., Tatusova, T. and Maglott, D.R. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*, **35**, D61-65.

89. Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. and Wheeler, D.L. (2007) GenBank. *Nucleic Acids Res*, **35**, D21-25.

90. Finn, R.D., Mistry, J., Schuster-Bockler, B., Griffiths-Jones, S., Hollich, V., Lassmann, T., Moxon, S., Marshall, M., Khanna, A., Durbin, R. *et al.* (2006) Pfam: clans, web tools and services. *Nucl. Acids Res.*, **34**, D247-251.

91. Letunic, I., Copley, R.R., Pils, B., Pinkert, S., Schultz, J. and Bork, P. (2006) SMART 5: domains in the context of genomes and networks. *Nucleic Acids Res*, **34**, D257-260.

92. Wetlaufer, D.B. (1973) Nucleation, rapid folding, and globular intrachain regions in proteins. *Proc Natl Acad Sci U S A*, **70**, 697-701.

93. Patthy, L. (1987) Detecting homology of distantly related proteins with consensus sequences. *J Mol Biol*, **198**, 567-577.

94. Sauer, R.T., Yocum, R.R., Doolittle, R.F., Lewis, M. and Pabo, C.O. (1982) Homology among DNA-binding proteins suggests use of a conserved super-secondary structure. *Nature*, **298**, 447-451.

95. Bork, P. (1991) Shuffled domains in extracellular proteins. *Febs Lett*, **286**, 47-54.

96. Bru, C., Courcelle, E., Carrere, S., Beausse, Y., Dalmar, S. and Kahn, D. (2005) The ProDom database of protein domain families: more emphasis on 3D. *Nucleic Acids Res*, **33**, D212-215.

97. Kouranov, A., Xie, L., de la Cruz, J., Chen, L., Westbrook, J., Bourne, P.E. and Berman, H.M. (2006) The RCSB PDB information portal for structural genomics. *Nucleic Acids Res*, **34**, D302-305.

98. Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.C., Estreicher, A., Gasteiger, E., Martin, M.J., Michoud, K., O'Donovan, C., Phan, I. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res*, **31**, 365-370.

99. Eddy, S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755-763.

100. Kim, H.S., Schell, M.A., Yu, Y., Ulrich, R.L., Sarria, S.H., Nierman, W.C. and DeShazer, D. (2005) Bacterial genome adaptation to niches: divergence of the potential virulence genes in three Burkholderia species of different survival strategies. *BMC genomics*, **6**, 174.

101. Wu, M., Ren, Q., Durkin, A.S., Daugherty, S.C., Brinkac, L.M., Dodson, R.J., Madupu, R., Sullivan, S.A., Kolonay, J.F., Haft, D.H. *et al.* (2005) Life in hot carbon monoxide: the complete genome sequence of Carboxydothermus hydrogenoformans Z-2901. *PLoS Genet*, **1**, e65.

102. Goldman, B.S., Nierman, W.C., Kaiser, D., Slater, S.C., Durkin, A.S., Eisen, J.A., Ronning, C.M., Barbazuk, W.B., Blanchard, M., Field, C. *et al.* (2006) Evolution of sensory complexity recorded in a myxobacterial genome. *Proc Natl Acad Sci U S A*, **103**, 15200-15205.

103. Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F. and Higgins, D.G. (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res*, **25**, 4876-4882.

104. Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, **32**, 1792-1797.

105. Galtier, N., Gouy, M. and Gautier, C. (1996) SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. *Comput Appl Biosci*, **12**, 543-548.

106. Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, **22**, 4673-4680.

107. Kumar, S., Tamura, K. and Nei, M. (2004) MEGA3: Integrated software for molecular evolutionary genetics analysis and sequence alignment. *Briefings in Bioinformatics*, **5**, 150-163.

108. Guindon, S. and Gascuel, O. (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic biology*, **52**, 696-704.

109. Cuff, J.A. and Barton, G.J. (2000) Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins*, **40**, 502-511.

110. Ulrich, L.E. and Zhulin, I.B. (2005) Four-helix bundle: a ubiquitous sensory module in prokaryotic signal transduction. *Bioinformatics*, **21**, iii45-48.

111. Jones, D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol*, **292**, 195-202.

112. Crooks, G.E., Hon, G., Chandonia, J.M. and Brenner, S.E. (2004) WebLogo: A sequence logo generator. *Genome Research*, **14**, 1188-1190.

113. Nomura, N., Sako, Y. and Uchida, A. (1998) Molecular characterization and postsplicing fate of three introns within the single rRNA operon of the hyperthermophilic archaeon Aeropyrum pernix K1. *J Bacteriol*, **180**, 3635-3643.

114. Holt, R.A., Subramanian, G.M., Halpern, A., Sutton, G.G., Charlab, R., Nusskern, D.R., Wincker, P., Clark, A.G., Ribeiro, J.M., Wides, R. *et al.* (2002) The genome sequence of the malaria mosquito Anopheles gambiae. *Science*, **298**, 129-149.

115. West, A.H. and Stock, A.M. (2001) Histidine kinases and response regulator proteins in two-component signaling systems. *Trends Biochem Sci*, **26**, 369-376.

116. McCarter, L.L. (2006) Regulation of flagella. *Current Opinion in Microbiology*, **9**, 180-186.

117. Journet, L., Hughes, K.T. and Cornelis, G.R. (2005) Type III secretion: a secretory pathway serving both motility and virulence (review). *Mol Membr Biol*, **22**, 41-50.

118. Thomas, N.A., Mueller, S., Klein, A. and Jarrell, K.F. (2002) Mutants in flaI and flaJ of the archaeon Methanococcus voltae are deficient in flagellum assembly. *Mol Microbiol*, **46**, 879-887.

119. Patenge, N., Berendes, A., Engelhardt, H., Schuster, S.C. and Oesterhelt, D. (2001) The fla gene cluster is involved in the biogenesis of flagella in Halobacterium salinarum. *Mol Microbiol*, **41**, 653-663.

120. Peabody, C.R., Chung, Y.J., Yen, M.R., Vidal-Ingigliardi, D., Pugsley, A.P. and Saier, M.H., Jr. (2003) Type II protein secretion and its relationship to bacterial type IV pili and archaeal flagella. *Microbiology*, **149**, 3051-3072.

121. Szurmant, H., Muff, T.J. and Ordal, G.W. (2004) Bacillus subtilis CheC and FliY are members of a novel class of CheY-P-hydrolyzing proteins in the chemotactic signal transduction cascade. *Journal of Biological Chemistry*, **279**, 21787-21792.

122. Planet, P.J., Kachlany, S.C., DeSalle, R. and Figurski, D.H. (2001) Phylogeny of genes for secretion NTPases: identification of the widespread tadA subfamily and development of a diagnostic key for gene classification. *Proc Natl Acad Sci U S A*, **98**, 2503-2508.

123. Bhaya, D., Bianco, N.R., Bryant, D. and Grossman, A. (2000) Type IV pilus biogenesis and motility in the cyanobacterium Synechocystis sp PCC6803. *Molecular Microbiology*, **37**, 941-951.

124. Park, H.S., Wolfgang, M. and Koomey, M. (2002) Modification of type IV pilus-associated epithelial cell adherence and multicellular behavior by the PilU protein of Neisseria gonorrhoeae. *Infect Immun*, **70**, 3891-3903.

125. Graupner, S., Weger, N., Sohni, M. and Wackernagel, W. (2001) Requirement of novel competence genes pilT and pilU of Pseudomonas stutzeri for natural transformation and suppression of pilT deficiency by a hexahistidine tag on the type IV pilus protein PilAI. *J Bacteriol*, **183**, 4694-4701.

126. Whitchurch, C.B. and Mattick, J.S. (1994) Characterization of a gene, pilU, required for twitching motility but not phage sensitivity in Pseudomonas aeruginosa. *Mol Microbiol*, **13**, 1079-1091.

127. Whitchurch, C.B., Hobbs, M., Livingston, S.P., Krishnapillai, V. and Mattick, J.S. (1991) Characterisation of a Pseudomonas aeruginosa twitching motility gene and evidence for a specialised protein export system widespread in eubacteria. *Gene*, **101**, 33-44.

128. Galperin, M.Y. (2006) Structural classification of bacterial response regulators: Diversity of output domains and domain combinations. *Journal of Bacteriology*, **188**, 4169-4182.

129. Tsoka, S. and Ouzounis, C.A. (2000) Prediction of protein interactions: metabolic enzymes are frequently involved in gene fusion. *Nat Genet*, **26**, 141-142.

130. Enright, A.J., Iliopoulos, I., Kyrpides, N.C. and Ouzounis, C.A. (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature*, **402**, 86-90.

131. Marcotte, E.M., Pellegrini, M., Ng, H.L., Rice, D.W., Yeates, T.O. and Eisenberg, D. (1999) Detecting protein function and protein-protein interactions from genome sequences. *Science*, **285**, 751-753.

132. Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D. and Yeates, T.O. (1999) Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 4285-4288.

133. Gaasterland, T. and Ragan, M.A. (1998) Microbial genescapes: phyletic and functional patterns of ORF distribution among prokaryotes. *Microbial & comparative genomics*, **3**, 199-217.

134. Blocker, A., Komoriya, K. and Aizawa, S. (2003) Type III secretion systems and bacterial flagella: Insights into their function from structural similarities. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 3027-3030.

135. Faguy, D.M., Jarrell, K.F., Kuzio, J. and Kalmokoff, M.L. (1994) Molecular analysis of archael flagellins: similarity to the type IV pilin-transport superfamily widespread in bacteria. *Canadian journal of microbiology*, **40**, 67-71.

136. Mattick, J.S. (2002) Type IV pili and twitching motility. *Annu Rev Microbiol*, **56**, 289-314.

137. Bardy, S.L., Ng, S.Y. and Jarrell, K.F. (2003) Prokaryotic motility structures. *Microbiology*, **149**, 295-304.

138. Skerker, J.M. and Berg, H.C. (2001) Direct observation of extension and retraction of type IV pili. *Proc Natl Acad Sci U S A*, **98**, 6901-6904.

139. Chiang, P., Habash, M. and Burrows, L.L. (2005) Disparate subcellular localization patterns of Pseudomonas aeruginosa Type IV pilus ATPases involved in twitching motility. *J Bacteriol*, **187**, 829-839.

140. Sun, H., Zusman, D.R. and Shi, W. (2000) Type IV pilus of Myxococcus xanthus is a motility apparatus controlled by the frz chemosensory system. *Curr Biol*, **10**, 1143-1146.

141. Reguera, G., McCarthy, K.D., Mehta, T., Nicoll, J.S., Tuominen, M.T. and Lovley, D.R. (2005) Extracellular electron transfer via microbial nanowires. *Nature*, **435**, 1098-1101.

142. Reguera, G., Pollina, R.B., Nicoll, J.S. and Lovley, D.R. (2007) Possible nonconductive role of Geobacter sulfurreducens pilus nanowires in biofilm formation. *J Bacteriol*, **189**, 2125-2127.

143. Fussenegger, M., Rudel, T., Barten, R., Ryll, R. and Meyer, T.F. (1997) Transformation competence and type-4 pilus biogenesis in Neisseria gonorrhoeae--a review. *Gene*, **192**, 125-134.

144. Dutta, R., Qin, L. and Inouye, M. (1999) Histidine kinases: diversity of domain organization. *Molecular Microbiology*, **34**, 633-640.

145. Terry, K., Williams, S.M., Connolly, L. and Ottemann, K.M. (2005) Chemotaxis plays multiple roles during Helicobacter pylori animal infection. *Infect Immun*, **73**, 803-811.

146. Rudolph, J. and Oesterhelt, D. (1996) Deletion analysis of the che operon in the archaeon Halobacterium salinarium. *J Mol Biol*, **258**, 548-554.

147. Scharf, B. and Schmitt, R. (2002) Sensory transduction to the flagellar motor of Sinorhizobium meliloti. *J Mol Microbiol Biotechnol*, **4**, 183-186.

148. Dons, L., Eriksson, E., Jin, Y., Rottenberg, M.E., Kristensson, K., Larsen, C.N., Bresciani, J. and Olsen, J.E. (2004) Role of flagellin and the two-component CheA/CheY system of Listeria monocytogenes in host cell invasion and virulence. *Infect Immun*, **72**, 3237-3244.

149. Sim, J.H., Shi, W. and Lux, R. (2005) Protein-protein interactions in the chernotaxis signalling pathway of Treponema denticola. *Microbiology-Sgm*, **151**, 1801-1807.

150. Yao, J. and Allen, C. (2006) Chemotaxis is required for virulence and competitive fitness of the bacterial wilt pathogen Ralstonia solanacearum. *J Bacteriol*, **188**, 3697-3708.

151. Ely, B., Gerardot, C.J., Fleming, D.L., Gomes, S.L., Frederikse, P. and Shapiro, L. (1986) General nonchemotactic mutants of Caulobacter crescentus. *Genetics*, **114**, 717-730.

152. Okumura, H., Nishiyama, S.I., Sasaki, A., Homma, M. and Kawagishi, I. (1998) Chemotactic adaptation is altered by changes in the carboxy-terminal sequence conserved among the major methyl-accepting chemoreceptors. *Journal of Bacteriology*, **180**, 1862-1868.

153. Hess, J.F., Oosawa, K., Kaplan, N. and Simon, M.I. (1988) Phosphorylation of three proteins in the signaling pathway of bacterial chemotaxis. *Cell*, **53**, 79-87.

154. Rosario, M.M., Fredrick, K.L., Ordal, G.W. and Helmann, J.D. (1994) Chemotaxis in Bacillus subtilis requires either of two functionally redundant CheW homologs. *J Bacteriol*, **176**, 2736-2739.

155. Motaleb, M.A., Miller, M.R., Li, C., Bakker, R.G., Goldstein, S.F., Silversmith, R.E., Bourret, R.B. and Charon, N.W. (2005) CheX is a phosphorylated CheY phosphatase essential for Borrelia burgdorferi chemotaxis. *J Bacteriol*, **187**, 7963-7969.

156. Wuichet, K. and Zhulin, I.B. (2003) Molecular evolution of sensory domains in cyanobacterial chemoreceptors. *Trends in Microbiology*, **11**, 200-203.

157. Surette, M.G., Levit, M., Liu, Y., Lukat, G., Ninfa, E.G., Ninfa, A. and Stock, J.B. (1996) Dimerization is required for the activity of the protein histidine kinase CheA that mediates signal transduction in bacterial chemotaxis. *Journal of Biological Chemistry*, **271**, 939-945.

158. Swanson, R.V., Schuster, S.C. and Simon, M.I. (1993) Expression of Chea Fragments Which Define Domains Encoding Kinase, Phosphotransfer, and Chey Binding Activities. *Biochemistry*, **32**, 7623-7629.

159. Boukhvalova, M.S., Dahlquist, F.W. and Stewart, R.C. (2002) CheW binding interactions with CheA and Tar - Importance for chemotaxis signaling in Escherichia coli. *Journal of Biological Chemistry*, **277**, 22251-22259.

160. Hess, J.F., Bourret, R.B. and Simon, M.I. (1988) Histidine phosphorylation and phosphoryl group transfer in bacterial chemotaxis. *Nature*, **336**, 139-143.

161. Parkinson, J.S. (1976) cheA, cheB, and cheC genes of Escherichia coli and their role in chemotaxis. *J Bacteriol*, **126**, 758-770.

162. Fuhrer, D.K. and Ordal, G.W. (1991) Bacillus subtilis CheN, a homolog of CheA, the central regulator of chemotaxis in Escherichia coli. *J Bacteriol*, **173**, 7443-7448.

163. Porter, S.L. and Armitage, J.P. (2004) Chemotaxis in Rhodobacter sphaeroides requires an atypical histidine protein kinase. *J Biol Chem*, **279**, 54573-54580.

164. Greck, M., Platzer, J., Sourjik, V. and Schmitt, R. (1995) Analysis of a chemotaxis operon in Rhizobium meliloti. *Mol Microbiol*, **15**, 989-1000.

165. Rudolph, J. and Oesterhelt, D. (1995) Chemotaxis and phototaxis require a CheA histidine kinase in the archaeon Halobacterium salinarium. *Embo J*, **14**, 667-673.

166. Flanary, P.L., Allen, R.D., Dons, L. and Kathariou, S. (1999) Insertional inactivation of the Listeria monocytogenes cheYA operon abolishes response to oxygen gradients and reduces the number of flagella. *Canadian journal of microbiology*, **45**, 646-652.

167. Lux, R., Sim, J.H., Tsai, J.P. and Shi, W. (2002) Construction and characterization of a cheA mutant of Treponema denticola. *Journal of Bacteriology*, **184**, 3130-3134.

168. Alley, M.R., Gomes, S.L., Alexander, W. and Shapiro, L. (1991) Genetic analysis of a temporally transcribed chemotaxis gene cluster in Caulobacter crescentus. *Genetics*, **129**, 333-341.

169. de Weert, S., Vermeiren, H., Mulders, I.H., Kuiper, I., Hendrickx, N., Bloemberg, G.V., Vanderleyden, J., De Mot, R. and Lugtenberg, B.J. (2002) Flagella-driven chemotaxis towards exudate components is an important trait for tomato root colonization by Pseudomonas fluorescens. *Mol Plant Microbe Interact*, **15**, 1173-1180.

170. Mirkin, B.G., Fenner, T.I., Galperin, M.Y. and Koonin, E.V. (2003) Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *Bmc Evol Biol*, **3**, -.

171. Yang, Z., Ma, X., Tong, L., Kaplan, H.B., Shimkets, L.J. and Shi, W. (2000) Myxococcus xanthus dif genes are required for biogenesis of cell surface fibrils essential for social gliding motility. *J Bacteriol*, **182**, 5793-5798.

172. Molofsky, A.B., Shetron-Rama, L.M. and Swanson, M.S. (2005) Components of the Legionella pneumophila flagellar regulon contribute to multiple virulence traits, including lysosome avoidance and macrophage death. *Infect Immun*, **73**, 5720-5734.

173. Badger, J.H., Hoover, T.R., Brun, Y.V., Weiner, R.M., Laub, M.T., Alexandre, G., Mrazek, J., Ren, Q., Paulsen, I.T., Nelson, K.E. *et al.* (2006) Comparative genomic evidence for a close relationship between the dimorphic prosthecate bacteria Hyphomonas neptunium and Caulobacter crescentus. *J Bacteriol*, **188**, 6841-6850.

174. Woolfit, M. and Bromham, L. (2003) Increased rates of sequence evolution in endosymbiotic bacteria and fungi with small effective population sizes. *Mol Biol Evol*, **20**, 1545-1555.

175. Koretke, K.K., Lupas, A.N., Warren, P.V., Rosenberg, M. and Brown, J.R. (2000) Evolution of two-component signal transduction. *Mol Biol Evol*, **17**, 1956-1970.

176. Chiang, P. and Burrows, L.L. (2003) Biofilm formation by hyperpiliated mutants of Pseudomonas aeruginosa. *J Bacteriol*, **185**, 2374-2378.

177. Ramboarina, S., Fernandes, P.J., Daniell, S., Islam, S., Simpson, P., Frankel, G., Booy, F., Donnenberg, M.S. and Matthews, S. (2005) Structure of the bundle-forming pilus from enteropathogenic Escherichia coli. *J Biol Chem*, **280**, 40252-40260.

178. Alm, E., Huang, K. and Arkin, A. (2006) The evolution of two-component systems in bacteria reveals different strategies for niche adaptation. *PLoS Comput Biol*, **2**, e143.

179. Schneider, T.D. and Stephens, R.M. (1990) Sequence Logos - a New Way to Display Consensus Sequences. *Nucleic Acids Research*, **18**, 6097-6100.

180. Darzins, A. (1995) The Pseudomonas aeruginosa pilK gene encodes a chemotactic methyltransferase (CheR) homologue that is translationally regulated. *Mol Microbiol*, **15**, 703-717.

181. Omelchenko, M.V., Wolf, Y.I., Gaidamakova, E.K., Matrosova, V.Y., Vasilenko, A., Zhai, M., Daly, M.J., Koonin, E.V. and Makarova, K.S. (2005) Comparative genomics of Thermus thermophilus and Deinococcus radiodurans: divergent routes of adaptation to thermophily and radiation resistance. *Bmc Evol Biol*, **5**, 57.

182. Lamb, J.R., Tugendreich, S. and Hieter, P. (1995) Tetratrico peptide repeat interactions: to TPR or not to TPR? *Trends Biochem Sci*, **20**, 257-259.

183. Karniol, B. and Vierstra, R.D. (2004) The HWE histidine kinases, a new family of bacterial two-component sensor kinases with potentially diverse roles in environmental signaling. *Journal of Bacteriology*, **186**, 445-453.

184. Shiomi, D., Zhulin, I.B., Homma, M. and Kawagishi, I. (2002) Dual recognition of the bacterial chemoreceptor by chemotaxis-specific domains of the CheR methyltransferase. *Journal of Biological Chemistry*, **277**, 42325-42333.

185. Taylor, B.L. and Zhulin, I.B. (1999) PAS domains: Internal sensors of oxygen, redox potential, and light. *Microbiology and Molecular Biology Reviews*, **63**, 479-506.

186. Zhulin, I.B., Nikolskaya, A.N. and Galperin, M.Y. (2003) Common extracellular sensory domains in transmembrane receptors for diverse signal transduction pathways in bacteria and archaea. *Journal of Bacteriology*, **185**, 285-294.

187. Hulko, M., Berndt, F., Gruber, M., Linder, J.U., Truffault, V., Schultz, A., Martin, J., Schultz, J.E., Lupas, A.N. and Coles, M. (2006) The HAMP domain structure implies helix rotation in transmembrane signaling. *Cell*, **126**, 929-940.

188. Aravind, L. and Ponting, C.P. (1997) The GAF domain: an evolutionary link between diverse phototransducing proteins. *Trends Biochem Sci*, **22**, 458-459.

189. Nowlin, D.M., Bollinger, J. and Hazelbauer, G.L. (1987) Sites of Covalent Modification in Trg, a Sensory Transducer of Escherichia-Coli. *Journal of Biological Chemistry*, **262**, 6039-6045.

190. Marina, A., Waldburger, C.D. and Hendrickson, W.A. (2005) Structure of the entire cytoplasmic portion of a sensor histidine-kinase protein. *Embo J*, **24**, 4247-4259.

191. Berg, H.C. and Purcell, E.M. (1977) Physics of chemoreception. *Biophys J*, **20**, 193-219.

192. Wu, J., Li, J., Li, G., Long, D.G. and Weis, R.M. (1996) The receptor binding site for the methyltransferase of bacterial chemotaxis is distinct from the sites of methylation. *Biochemistry*, **35**, 4984-4993.

193. Jimenez-Pearson, M.-A., Delany, I., Scarlato, V. and Beier, D. (2005) Phosphate flow in the chemotactic response system of Helicobacter pylori. *Microbiology*, **151**, 3299-3311.

194. Kirby, J.R., Kristich, C.J., Saulmon, M.M., Zimmer, M.A., Garrity, L.F., Zhulin, I.B. and Ordal, G.W. (2001) CheC is related to the family of flagellar switch proteins and acts independently from CheD to control chemotaxis in Bacillus subtilis. *Molecular Microbiology*, **42**, 573-585.

195. Kehry, M.R., Bond, M.W., Hunkapiller, M.W. and Dahlquist, F.W. (1983) Enzymatic Deamidation of Methyl-Accepting Chemotaxis Proteins in Escherichia-Coli Catalyzed by the Cheb Gene-Product. *P Natl Acad Sci-Biol*, **80**, 3599-3603.

196. Barnakov, A.N., Barnakova, L.A. and Hazelbauer, G.L. (2001) Location of the receptor-interaction site on CheB, the methylesterase response regulator of bacterial chemotaxis. *Journal of Biological Chemistry*, **276**, 32984-32989.

197. Terry, K., Go, A.C. and Ottemann, K.M. (2006) Proteomic mapping of a suppressor of non-chemotactic cheW mutants reveals that Helicobacter pylori contains a new chemotaxis protein. *Molecular Microbiology*, **61**, 871-882.

198. Pittman, M.S., Goodwin, M. and Kelly, D.J. (2001) Chemotaxis in the human gastric pathogen Helicobacter pylori: different roles for CheW and the three CheV paralogues, and evidence for CheV2 phosphorylation. *Microbiology*, **147**, 2493-2504.

199. Marchant, J., Wren, B. and Ketley, J. (2002) Exploiting genome sequence: predictions for mechanisms of Campylobacter chemotaxis. *Trends in Microbiology*, **10**, 155-159.

200. Cantwell, B.J., Draheim, R.R., Weart, R.B., Nguyen, C., Stewart, R.C. and Manson, M.D. (2003) CheZ phosphatase localizes to chemoreceptor patches via CheA-short. *Journal of Bacteriology*, **185**, 2354-2361.

201. McEvoy, M.M., Bren, A., Eisenbach, M. and Dahlquist, F.W. (1999) Identification of the binding interfaces on CheY for two of its targets, the phosphatase CheZ and the flagellar switch protein fliM. *J Mol Biol*, **289**, 1423-1433.

202. Zhu, X., Volz, K. and Matsumura, P. (1997) The CheZ-binding surface of CheY overlaps the CheA- and FliM-binding surfaces. *J Biol Chem*, **272**, 23758-23764.

203. Lee, S.Y., Cho, H.S., Pelton, J.G., Yan, D., Henderson, R.K., King, D.S., Huang, L., Kustu, S., Berry, E.A. and Wemmer, D.E. (2001) Crystal structure of an activated response regulator bound to its target. *Nat Struct Biol*, **8**, 52-56.

204. Black, W.P. and Yang, Z. (2004) Myxococcus xanthus chemotaxis homologs DifD and DifG negatively regulate fibril polysaccharide production. *J Bacteriol*, **186**, 1001-1008.

205. Boukhvalova, M., VanBruggen, R. and Stewart, R.C. (2002) CheA kinase and chemoreceptor interaction surfaces on CheW. *Journal of Biological Chemistry*, **277**, 23596-23603.

206. Anantharaman, V. and Aravind, L. (2000) Cache - a signaling domain common to animal Ca(2+)-channel subunits and a class of prokaryotic chemotaxis receptors. *Trends Biochem Sci*, **25**, 535-537.

207. Makarova, K.S., Koonin, E.V., Haselkorn, R. and Galperin, M.Y. (2006) Cyanobacterial response regulator PatA contains a conserved N-terminal domain (PATAN) with an alpha-helical insertion. *Bioinformatics*, **22**, 1297-1301.

208. Liang, J., Scappino, L. and Haselkorn, R. (1992) The patA gene product, which contains a region similar to CheY of Escherichia coli, controls heterocyst pattern formation in the cyanobacterium Anabaena 7120. *Proc Natl Acad Sci U S A*, **89**, 5655-5659.

209. Ueda, K., Yamashita, A., Ishikawa, J., Shimada, M., Watsuji, T.O., Morimura, K., Ikeda, H., Hattori, M. and Beppu, T. (2004) Genome sequence of Symbiobacterium thermophilum, an uncultivable bacterium that depends on microbial commensalism. *Nucleic Acids Res*, **32**, 4937-4944.

210. Trivedi, V.D. and Spudich, J.L. (2003) Photostimulation of a sensory rhodopsin II/HtrII/Tsr fusion chimera activates CheA-autophosphorylation and CheY-phosphotransfer in vitro. *Biochemistry*, **42**, 13887-13892.

211. Ciccarelli, F.D., Doerks, T., von Mering, C., Creevey, C.J., Snel, B. and Bork, P. (2006) Toward automatic reconstruction of a highly resolved tree of life. *Science*, **311**, 1283-1287.

212. Lu, X. and Li, Y. (2001) Simulation of the evolution of genomic complexity. *Bio Systems*, **61**, 83-94.

213. Roth, G., Nishikawa, K.C. and Wake, D.B. (1997) Genome size, secondary simplification, and the evolution of the brain in salamanders. *Brain, behavior and evolution*, **50**, 50-59.

214. Ren, C.P., Beatson, S.A., Parkhill, J. and Pallen, M.J. (2005) The Flag-2 locus, an ancestral gene cluster, is potentially associated with a novel flagellar system from Escherichia coli. *Journal of Bacteriology*, **187**, 1430-1440.

215. Sar, N., McCarter, L., Simon, M. and Silverman, M. (1990) Chemotactic control of the two flagellar systems of Vibrio parahaemolyticus. *J Bacteriol*, **172**, 334-341.

216. Kojima, M., Kubo, R., Yakushi, T., Homma, M. and Kawagishi, I. (2007) The bidirectional polar and unidirectional lateral flagellar motors of Vibrio alginolyticus are controlled by a single CheY species. *Mol Microbiol*, **64**, 57-67.

217. Hyakutake, A., Homma, M., Austin, M.J., Boin, M.A., Hase, C.C. and Kawagishi, I. (2005) Only one of the five CheY homologs in Vibrio cholerae directly switches flagellar rotation. *J Bacteriol*, **187**, 8403-8410.

218. Poggio, S., Abreu-Goodger, C., Fabela, S., Osorio, A., Dreyfus, G., Vinuesa, P. and Camarena, L. (2007) A Complete Set of Flagellar Genes Acquired by Horizontal Transfer Coexists with the Endogenous Flagellar System in Rhodobacter sphaeroides. *J Bacteriol*, **189**, 3208-3216.

219. Anderson, F.E. and Swofford, D.L. (2004) Should we be worried about long-branch attraction in real data sets? Investigations using metazoan 18S rDNA. *Mol Phylogenet Evol*, **33**, 440-451.

220. Dusenbery, D.B. (1998) Spatial sensing of stimulus gradients can be superior to temporal sensing for free-swimming bacteria. *Biophys J*, **74**, 2272-2277.

221. Cavalier-Smith, T. (2006) Rooting the tree of life by transition analyses. *Biology direct*, **1**, 19.

222. Gupta, R.S. (2000) The natural evolutionary relationships among prokaryotes. *Critical reviews in microbiology*, **26**, 111-131.

223. Liu, R. and Ochman, H. (2007) Stepwise formation of the bacterial flagellar system. *Proc Natl Acad Sci U S A*, **104**, 7116-7121.

224. Mushegian, A.R. and Koonin, E.V. (1996) A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proc Natl Acad Sci U S A*, **93**, 10268-10273.

225. Kunin, V. and Ouzounis, C.A. (2003) The balance of driving forces during genome evolution in prokaryotes. *Genome Res*, **13**, 1589-1594.

226. Snel, B., Bork, P. and Huynen, M.A. (2002) Genomes in flux: the evolution of archaeal and proteobacterial gene content. *Genome Res*, **12**, 17-25.

227. Koonin, E.V. (2000) How many genes can make a cell: the minimal-gene-set concept. *Annual review of genomics and human genetics*, **1**, 99-116.

228. Zhulin, I.B. (2001) The superfamily of chemotaxis transducers: from physiology to genomics and back. *Advances in Microbial Physiology, Vol 41*, **45**, 157-198.

229. Zhulin, I.B. (2000) A novel phototaxis receptor hidden in the cyanobacterial genome. *J Mol Microbiol Biotechnol*, **2**, 491-493.

230. Shu, C.J., Ulrich, L.E. and Zhulin, I.B. (2003) The NIT domain: a predicted nitrate-responsive module in bacterial sensory receptors. *Trends Biochem Sci*, **28**, 121-124.

231. LeMoual, H. and Koshland, D.E. (1996) Molecular evolution of the C-terminal cytoplasmic domain of a superfamily of bacterial receptors involved in taxis. *Journal of Molecular Biology*, **261**, 568-585.

232. Wolf, Y.I., Rogozin, I.B., Grishin, N.V., Tatusov, R.L. and Koonin, E.V. (2001) Genome trees constructed using five different approaches suggest new major bacterial clades. *Bmc Evol Biol*, **1**, 8.

233. Itoh, T., Takemoto, K., Mori, H. and Gojobori, T. (1999) Evolutionary instability of operon structures disclosed by sequence comparisons of complete microbial genomes. *Molecular Biology and Evolution*, **16**, 332-346.

234. Yoshihara, S., Geng, X. and Ikeuchi, M. (2002) pilG Gene cluster and split pilL genes involved in pilus biogenesis, motility and genetic transformation in the cyanobacterium Synechocystis sp. PCC 6803. *Plant & cell physiology*, **43**, 513-521.

235. Kelley, L.A., MacCallum, R.M. and Sternberg, M.J. (2000) Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J Mol Biol*, **299**, 499-520.

236. Yoshihara, S., Suzuki, F., Fujita, H., Geng, X.X. and Ikeuchi, M. (2000) Novel putative photoreceptor and regulatory genes Required for the positive phototactic movement of the unicellular motile cyanobacterium Synechocystis sp. PCC 6803. *Plant & cell physiology*, **41**, 1299-1304.

237. Alexandre, G. and Zhulin, I.B. (2003) Different evolutionary constraints on chemotaxis proteins CheW and CheY revealed by heterologous expression studies and protein sequence analysis. *J Bacteriol*, **185**, 544-552.

238. Falke, J.J. and Hazelbauer, G.L. (2001) Transmembrane signaling in bacterial chemoreceptors. *Trends in Biochemical Sciences*, **26**, 257-265.

239. Koonin, E.V., Wolf, Y.I. and Karev, G.P. (2002) The structure of the protein universe and genome evolution. *Nature*, **420**, 218-223.

240. Fariselli, P., Pazos, F., Valencia, A. and Casadio, R. (2002) Prediction of protein-protein interaction sites in heterocomplexes with neural networks. *European Journal of Biochemistry*, **269**, 1356-1361.

241. Jones, S. and Thornton, J.M. (1997) Prediction of protein-protein interaction sites using patch analysis. *J Mol Biol*, **272**, 133-143.

242. Casari, G., Sander, C. and Valencia, A. (1995) A Method to Predict Functional Residues in Proteins. *Nature Structural Biology*, **2**, 171-178.

243. Gallet, X., Charloteaux, B., Thomas, A. and Brasseur, R. (2000) A fast method to predict protein interaction sites from sequences. *Journal of Molecular Biology*, **302**, 917-926.

244. Kini, R.M. and Evans, H.J. (1996) Prediction of potential protein-protein interaction sites from amino acid sequence. Identification of a fibrin polymerization site. *Febs Lett*, **385**, 81-86.

245. Ofran, Y. and Rost, B. (2003) Predicted protein-protein interaction sites from local sequence information. *Febs Lett*, **544**, 236-239.

246. Wang, B., Chen, P., Huang, D.S., Li, J.J., Lok, T.M. and Lyu, M.R. (2006) Predicting protein interaction sites from residue spatial sequence profile and evolution rate. *Febs Letters*, **580**, 380-384.

247. Yan, C., Dobbs, D. and Honavar, V. (2004) A two-stage classifier for identification of protein-protein interface residues. *Bioinformatics*, **20 Suppl 1**, I371-I378.

248. Kyte, J. and Doolittle, R.F. (1982) A simple method for displaying the hydropathic character of a protein. *J Mol Biol*, **157**, 105-132.

249. Lo Conte, L., Chothia, C. and Janin, J. (1999) The atomic structure of protein-protein recognition sites. *Journal of Molecular Biology*, **285**, 2177-2198.

250. Blat, Y. and Eisenbach, M. (1996) Conserved C-terminus of the phosphatase CheZ is a binding domain for the chemotactic response regulator CheY. *Biochemistry*, **35**, 5679-5683.

251. Sanna, M.G., Swanson, R.V., Bourret, R.B. and Simon, M.I. (1995) Mutations in the chemotactic response regulator, CheY, that confer resistance to the phosphatase activity of CheZ. *Mol Microbiol*, **15**, 1069-1079.

252. Silversmith, R.E. (2005) High mobility of carboxyl-terminal region of bacterial chemotaxis phosphatase CheZ is diminished upon binding divalent cation or CheY-P substrate. *Biochemistry*, **44**, 7768-7776.

253. Silversmith, R.E., Guanga, G.P., Betts, L., Chu, C., Zhao, R. and Bourret, R.B. (2003) CheZ-mediated dephosphorylation of the Escherichia coli chemotaxis response regulator CheY: role for CheY glutamate 89. *J Bacteriol*, **185**, 1495-1502.

254. Zhu, X., Amsler, C.D., Volz, K. and Matsumura, P. (1996) Tyrosine 106 of CheY plays an important role in chemotaxis signal transduction in Escherichia coli. *J Bacteriol*, **178**, 4208-4215.

255. Dekker, J.P., Fodor, A., Aldrich, R.W. and Yellen, G. (2004) A perturbation-based method for calculating explicit likelihood of evolutionary co-variance in multiple sequence alignments. *Bioinformatics*, **20**, 1565-1572.